Weakly-Supervised Learning of Category-Specific 3D Object Shapes

Junwei Han, *Senior Member, IEEE*, Yang Yang, Dingwen Zhang, Dong Huang, Dong Xu, *Fellow, IEEE*, and Fernando De La Torre

Abstract—Category-specific 3D object shape models have greatly boosted the recent advances in object detection, recognition and segmentation. However, even the most advanced approach for learning 3D object shapes still requires heavy manual annotations on large-scale 2D images. Such annotations include object categories, object keypoints, and figure-ground segmentation for the instances in each image. In particular, annotating figure-ground segmentation is unbearably labor-intensive and time-consuming. To address this problem, this paper devotes to learn category-specific 3D shape models under weak supervision, where only object categories and keypoints are required to be manually annotated on the training 2D images. By exploring the underlying relationship between two tasks: object segmentation and category-specific 3D shape reconstruction, we propose a novel weakly-supervised learning framework to jointly address these two tasks and combine them to boost the final performance of the learned 3D shape models. Moreover, learning without using figure-ground segmentation leads to ambiguous solutions. To this end, we develop the confidence weighting schemes in the viewpoint estimation and 3D shape learning procedure. These schemes effectively reduce the confusion caused by the noisy data and thus increase the chances for recovering more reliable 3D object shapes. Comprehensive experiments on the challenging PASCAL VOC benchmark show that our framework achieves comparable performance with the state-of-the-art methods that use expensive manual segmentation-level annotations. In addition, our experiments also demonstrate that our 3D shape models improve object segmentation performance.

Index Terms—3D shape reconstruction, common object segmentation, viewpoint estimation

1 Introduction

Recent works in computer vision community have achieved great success in object detection [1], [2], [3], [4], [5], event analysis, and object segmentation [6], [7], [8]. it is a challenging task to construct a rich internal representation of objects, such as the depth information and 3D pose. To address this problem, a two-step approach is widely adopted by the existing approaches. First, object-specific 3D shape models are obtained either by manual annotation or learning from the training data; Then, the obtained 3D shape models are used to align the objects in the test image and estimate their 3D depth and pose. In this paper, we mainly focus on the first step, i.e., how to learn the 3D shape models from the 2D image.

- *J. Han and Y. Yang are with the School of Automation, Northwestern Polytechnical University, Xi'an 710129, China. E-mail: junweihan2010@gmail.com, tp030ny@mail.nwpu.edu.cn.*
- *D. Zhang is with the School of Automation, Northwestern Polytechnical University, Xi'an 710129, China, and also with the School of Mechano-electronic Engineering, Xidian University, Xi'an 710126, China. E-mail: zhangdingwen2006yyy@gmail.com.*
- *D. Huang and F. Torre are with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213 USA. E-mail: dghuang@andrew.cmu.edu, ftorre@cs.cmu.edu.*
- *D. Xu is with the School of Electrical and Information Engineering, University of Sydney, Camperdown, NSW 2006, Australia. E-mail: dong.xu@sydney.edu.au.*

Manuscript received 2 Oct. 2018; revised 13 June 2019; accepted 7 Oct. 2019. Date of publication 25 Oct. 2019; date of current version 4 Mar. 2021. (Corresponding author: Dingwen Zhang). Recommended for acceptance by A. Geiger. Digital Object Identifier no. 10.1109/TPAMI.2019.2949562

The conventional approaches for acquiring 3D shape models follow a time-consuming and expensive process. They usually require manual design of the 3D shape models by using special 3D scan equipments with controlled imaging environments. To overcome such limitation, an alternative approach, which is similarly used in some recent works [9], [10], [11], is to build 3D object shape models by only using 2D images from the publicly available benchmark datasets, e.g., PASCAL VOC and ImageNet. Although these approaches are more convenient and less expensive than the conventional approaches, they still require labor-intensive and time-consuming manual annotation of 2D images, which includes annotation of object class labels, keypoints, and figure-ground segmentation masks.

Along this line of research, this paper makes a further effort to learn 3D shape models by only using weakly-labeled 2D images, and this can be also formulated as a weakly supervised 3D object reconstruction process, in which the 3D shape models are learnt from the 3D object reconstruction process of the weakly labelled individual training images. Here, weakly-labeled 2D images are the images only annotated with the corresponding object categories and a small number of keypoints of each object instance. Whereas the most time-consuming 2D manual annotation process for obtaining the pixel-level figure-ground segmentation mask of each object instance is not required any more. Obviously, learning category-specific 3D shape models in such way would be highly valuable as it bridges the 3D visual world with limited data and the 2D visual world with plenty of data. However, we need to address many challenges for

Fig. 1. Examples from the PASCAL VOC dataset. It is easier to learn 3D models by using the instances with less complex shapes and in relatively clean backgrounds. The learning process may become difficult when the instances are with complex shapes and in cluttered image backgrounds.

developing such a learning framework: (1) The figure-ground segmentation masks of the object instances play a key role in matching 3D object shapes with pixels in 2D images, but it is difficult for them to be inferred in complex scenarios. (2) Without strong supervision information from the manually labelled pixel-level masks, weakly-supervised learning methods face serious ambiguity issues when segmenting the 2D object masks and reconstructing the 3D object shapes. (3) Estimating 2D masks from weakly-labeled images is a necessary task that may be affected by significant image noise, and the errors in the estimated 2D masks will be significantly amplified in the final 3D shape models.

To address the first challenge, we introduce a hidden task by explicitly inferring the figure-ground segmentation mask of each training instance. Note this is different from the conventional weakly supervised learning frameworks that model 2D information as hidden variables of the objective function for 3D shape models. We take this approach because the structural space of the figure-ground segmentation masks is so large and it is hard to find the optimal solutions to hidden variables. In our approach, explicitly inferring segmentation masks produces a more reliable solution for learning the 3D shape models. Based on this idea, we introduce a Common Object Segmentation (COS) task into the learning framework. The goal of COS is to segment the common objects that appear in the images from the same object category. The segmentation masks obtained by COS can readily provide helpful supervision to guide the learning process of 3D object shapes.

Although COS can provide reliable segmentation masks, we still cannot build satisfactory 3D shape models by directly using the estimated segmentation masks, due to the aforementioned second challenge. Under the weakly-supervised scenario, neither the COS task nor the 3D object reconstruction task can be well-solved individually. Fortunately, we observe these two tasks can help each other to boost the performance of both tasks. On one hand, the figure-ground object masks generated by COS can help provide informative bottom-up shape cues for 3D object reconstruction. On the other hand, the 3D shape models from 3D object reconstruction can provide helpful yet under-explored top-down priors for COS. Based on this observation, we propose to solve COS and 3D object reconstruction alternatingly. When solving the COS task, we use the 3D shape models learnt from 3D object reconstruction to construct a prior-knowledge based term in the objective function of COS. When solving the 3D object reconstruction task, we apply the object segmentation masks obtained from the COS task to constrain the learning process of 3D shape models.

To address the third challenge, we propose a novel learning framework to learn the weakly superivsed category-specific 3D shape models with a confidence weighting scheme. The key observation is that, instances with simpler shapes and cleaner background can improve the learning process as it is easier to segment and reconstruct them (see the images in the left subfigure of Fig. 1). Meanwhile, instances with complex shapes and in the cluttered background usually confuse the learner as these instances will inevitably introduce noise into the learning process (see the images in the right subfigure of Fig. 1). Based on the above discussion, we propose a novel 3D shape reconstruction algorithm that computes the confidence of training samples at each learning iteration based on the currently learnt 3D shape models and the previously estimated segmentation masks. Then, the 3D shape models always learn more from instances with higher confidence. This confidence weighting scheme effectively reduces the confusion caused by noisy data and significantly boosts the final performance.

Fig. 2 illustrates the flow chart of our approach. We first collect the category-specific instances according to the object categories. Then, given the annotated keypoints, we use an improved Structure from Motion (SfM) method to estimate the camera viewpoint parameters for each of the training instances (See Fig. 2a). Specifically, we propose a new learning strategy to guide the viewpoint estimation process. We empirically demonstrate that this strategy can avoid bad

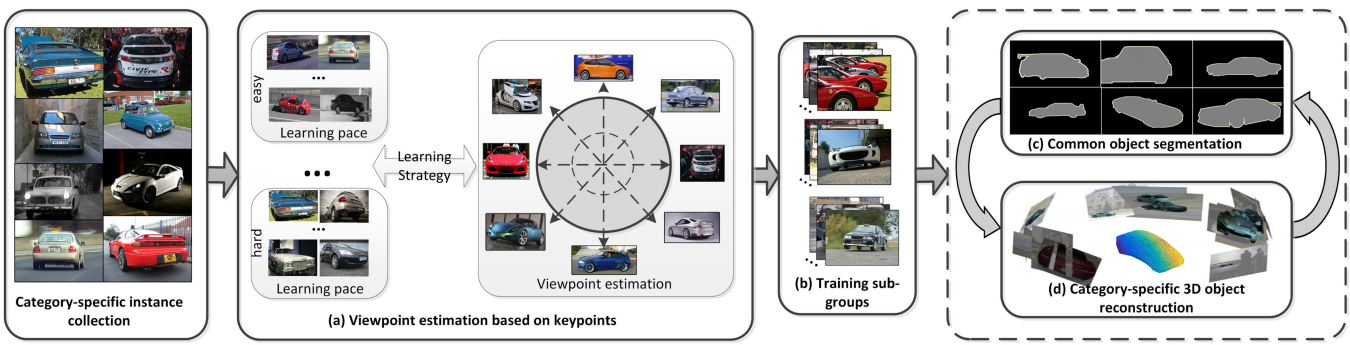


Fig. 2. The learning process (i.e., subfigures a-d) of the newly proposed 3D shape model learning framework. The subfigures (a) and (b) show the viewpoint estimation process (in Section 3.1) and the sub-group generation process (in Section 3.2). The subfigures (c) and (d) show the proposed iterative learning process for common object segmentation (in Section 3.4) and the 3D object reconstruction (in Section 3.3).

local minimum and achieve better viewpoint estimation results. Next, the category-specific instance collection is decomposed into subgroups (See Fig. 2b) and the initial coarse segmentation masks are obtained by using an existing object co-segmentation method [12]. Afterwards, we implement the joint optimization framework to infer the segmentation masks of each object instance and the 3D shape models of the specific object category. In this optimization framework, we first design a confidence weighting-based 3D object reconstruction algorithm to learn 3D shape models based on the estimated view points and the estimated 2D segmentation masks. Then, the learnt 3D shape models are used to provide the top-down prior for Common Object Segmentation. We alternatingly conduct the 3D object reconstruction process (See Fig. 2d) and the COS optimization process (See Fig. 2c) until the whole learning procedure converges. Finally, the segmentation masks and the category-specific 3D shape models are jointly learnt by the proposed weakly supervised learning framework.

In summary, this work mainly has fourfold contribution:

- We made one of the earliest effort to learn category-specific 3D object models only from weakly annotated 2D images. It can largely save the time and costs for manually labeling the figure-ground segmentation masks that are required by the existing methods, e.g., [9], [13].
- By revealing the underlying relationship between the common object segmentation task and the 3D object reconstruction task, we propose a novel framework to jointly solve the problems in these tasks and improve both of them.
- To cope with the noisy data during the joint learning process, we propose a robust viewpoint estimation method and and a confidence weighting-based 3D object reconstruction algorithm.
- Comprehensive experiments are implemented on the PASCAL VOC benchmark and the effectiveness of the proposed framework is demonstrated based on both the recovered category-specific 3D shape models and the estimated 2D segmentation masks.

The work in this paper is a substantial extension of our preliminary conference work in [14]. Compared with [14], the major differences in this paper are summarized below: 1) We improve the conventional Structure from Motion method (i.e., EM-PPCA [15]), by employing a new learning strategy for viewpoint estimation, which can effectively avoid bad local minimum and achieve better viewpoint estimation results. 2) We propose a novel 3D object reconstruction algorithm for learning the category-specific object shape models. Specifically, we additionally introduce a confidence weighting scheme into the learning process rather than equally treating all the training samples. In this way, the newly proposed 3D object reconstruction algorithm can address the learning ambiguity issue during its learning process and thus obtain more reliable 3D shape models. 3) More experiments are conducted with detailed analysis, including comparison for the 3D object reconstruction task, the common object segmentation task and the viewpoint estimation task.

2 Related Works

Learning 3D shape models is a long-standing and challenging research area in computer vision. The early research is under fixed-model-based paradigm, which focuses on exploiting 3D prior information by CAD models [16], [17] or special equipment (3D scanners) [18]. These approaches cannot work well when we are provided with in-the-wild training data containing unknown object categories and cluttered image background. For example, Choy et al. [19] employed the ground truth shape (the 3D CAD model) as 3D shape prior to provide supervision for training the end-to-end network. Rezende et al. [20] proposed to learn a generative model of 3D structures from volumetric data or multi-view images of individual object. Different from these methods, our goal is to reconstruct 3D shape prior for a category-level object from the weakly labeled dataset.

For category-level modeling, the works in [21], [22], [23], [24] learned morphable models of an object category by using 3D shapes of multiple object instances. These works are limited to simple objects as they assume both shapes and appearances of objects from the same category can be modeled by low-dimensional manifold/subspaces. Among these works, Zia et al. [22] proposed 3D geometric object category representations for object recognition from a single image. This approach recovers geometrically far more accurate object hypotheses rather than just bounding boxes. Specifically, the object hypotheses obtained in [22] include continuous estimation of object pose and 3D wireframes with relative 3D positions of object parts. By treating the shape of an object instance as a warped version of the mean shape of the category, Bao et al. [23] learned 3D prior (mean shape and a set of anchor points) from 3D scans and images of objects from various view points. Similarly, Dame et al. [24] combined image and 3D scans to improve the robustness of the classical 3D object reconstruction approaches. In contrast to the 3D data-based approaches, the previous data-driven methods [25], [26] explored traditional NRSfM [27] on the simple object categories, which reconstructs category-level 3D models directly from unordered images.

The recent works in [10], [11] learned category-level 3D shape models for complex object categories via the data-driven fashion which are closely related to our approach. Given the labeled training 2D data, the works in [10] estimated the viewpoints for each training instance based on NRSfM, and then recovered the 3D models for each category by deforming the surface mesh iteratively until convergence. However, these works heavily rely on the manually annotated 2D object masks. Compared with these approaches, our work makes the earliest step towards learning category-level 3D models from weakly labeled data. Our approach iteratively solves the common object segmentation task and category-specific 3D shape reconstruction task, and improves the performance of both tasks.

3 The Proposed Approach

Our approach learns the category-specific 3D shape model from the instances labelled with a specific object category, which consists of four main steps: (1) Estimating the view point of each training instance; (2) Clustering the instances according to their orientations and appearances and

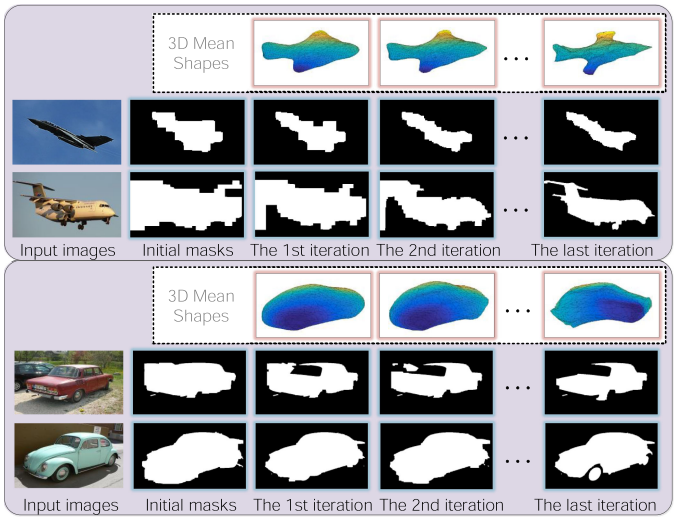


Fig. 3. Two examples (from the car and aeroplane categories) to show the evolvement of the segmentation masks and the 3D mean shapes when the number of iterations increases. Both the segmentation masks and 3D mean shapes are very coarse at the beginning, while we observe fine details after using our iterative learning approach.

initializing the segmentation mask for each training instance; (3) Learning category-specific 3D shape models by using the segmentation masks; (4) Segmenting objects by using category-specific 3D shape models. Step (3) and step (4) are performed alternatingly until convergence. By performing the aforementioned steps within our proposed learning framework, the 3D mean shape and the segmentation masks of each object of interest can be gradually improved when the number of iterations increases (see Fig. 3).

3.1 Viewpoint Estimation Based on Robust NRSfM

The object viewpoint is represented by camera parameters that project 3D shapes of an object to their 2D mask in images (See Fig. 2a). Given the training images and manually annotated 2D keypoints, the first step is to estimate the 3D coordinates of the keypoints. We propose a robust Non-Rigid Structure From Motion (Robust NRSfM) algorithm to reconstruct 3D coordinates of the keypoints and compute the camera parameters. As discussed in [9], the NRSfM method is a commonly used approach for camera viewpoint estimation from sparse correspondence as intra-class variation may degrade the performance significantly if it is not modeled explicitly.

The framework of NRSfM is performed on all object instances from the same category [15], [27]. We first crop the pixels of object instances from the images based on the rectangles enclosing the annotated 2D keypoints. Given K_p keypoints $\mathbf{P}_n \in \mathbb{R}^{2\times K_p}$ for the nth training instance $n \in \{1, 2, ..., N\}$. The NRSfM [9] algorithm maximizes the likelihood of the following formulation to approximate the 3D keypoint locations $\mathbf{W}_n \in \mathbb{R}^{3\times K_p}$ and the camera parameters $(c_n, \mathbf{R}_n, \mathbf{t}_n)$:

$$\begin{aligned} &\mathbf{P}_n = c_n\mathbf{R}_n\mathbf{W}_n + \mathbf{t}_n\mathbf{1}^T + \mathbf{H}_n, \\ &W_n = \overline{\mathbf{W}} + \sum\nolimits_d \mathbf{U}_d z_{n,d}, \\ &z_{n,d} \sim \mathcal{N}(0,1), H_{n,\iota} \sim \mathcal{N}(0, \sigma^2\mathbf{I}), \\ &d \in \{1, ..., D\}, \iota \in \{1, ..., K_p\}, \\ &s.t.\ \ \mathbf{R}_n\mathbf{R}_n^T = \mathbf{I}, \end{aligned} \tag{1}$$

where $\mathbf{I}$ denotes the identity matrix. $\mathbf{H}_n = [H_{n,1}, H_{n,2}, ..., H_{n,K_p}] \in \mathbb{R}^{2\times K_p}$ denotes the white noise matrix. The camera parameters include the rotation matrix $\mathbf{R}_n \in \mathbb{R}^{2\times 3}$ (orthographic projection), the scale c_n and the 2D translation $\mathbf{t}_n \in \mathbb{R}^{2\times 1}$. $\mathbf{1} \in \mathbb{R}^{K_p\times 1}$. The 3D keypoint $\mathbf{W}_n$ is parameterized based on a Gaussian distribution with a mean shape $\overline{\mathbf{W}}$, D deformation bases $\mathbf{U} = \{\mathbf{U}_1, \mathbf{U}_2, ..., \mathbf{U}_D\}$ and the deformation parameters $\mathbf{z}_n = \{z_{n,1}, z_{n,2}, ..., z_{n,D}\}$. We adopt the EM-PPCA algorithm [15] and the self-paced learning strategy [28], [29] to maximize the likelihood of Eq. 1. The self-paced learning theory [17] mainly considers the sample easiness to gradually learn from easy training samples to the complex ones, this learning strategy can avoid a bad local minimum when handling non-convex objectives.

We follow [15] to model the objective function as the Gaussian distribution of $\mathbf{P}_n$ by marginalizing the hidden variable $\mathbf{z}_n$, and then utilize the EM algorithm to optimize the parameters c_n, $\mathbf{R}_n$, $\mathbf{t}_n$, $\overline{\mathbf{W}}$, $\mathbf{U}$ and σ^2. Specifically, in the E-step, we estimate the posterior probability $Q(\mathbf{z}_n)$ according to the parameters from the last M-step. In the M-step, we estimate the parameters $\Psi = (c_n, \mathbf{R}_n, \mathbf{t}_n, \overline{\mathbf{W}}, \mathbf{U})$ and σ^2 according to $Q(\mathbf{z}_n)$. As such a latent variable model might strongly depend on initialization and would be easily trapped in to a bad local minimum, we introduce the self-paced regularizer [28] into the objective function of the M-step. Thus, we arrive at our objective function of the M-step as follows:[1]

$$\begin{aligned} \underset{\Psi,\sigma^2,\varepsilon\in\{0,1\}^N}{min} &\sum\nolimits_n \varepsilon_n \mathbb{E}_{\mathbf{z}_n\sim Q(\mathbf{z}_n)}\left[\log\left(\frac{1}{\rho(\mathbf{p}_n, \mathbf{z}_n|\Psi, \sigma^2)}\right)\right] \\ &- \eta\sum\nolimits_n \varepsilon_n, \end{aligned} \tag{2}$$

The selecting weights $\{\varepsilon_1, ..., \varepsilon_N\}$ determines which samples will be selected during the learning process. Following the self-paced learning theory [28], if $\mathbb{E}_{\mathbf{z}_n\sim Q(\mathbf{z}_n)}\log[(\frac{1}{\rho(\mathbf{P}_n,\mathbf{z}_n|\Psi,\sigma^2)})] \leq \eta$, then we have $\varepsilon_n = 1$, which means the nth training instance is suitable to be employed for the current training stage. On the other hand, if $\mathbb{E}_{\mathbf{z}_n\sim Q(\mathbf{z}_n)}[\log(\frac{1}{\rho(\mathbf{P}_n,\mathbf{z}_n|\Psi,\sigma^2)})] > \eta$, then we have $\varepsilon_n = 0$. The parameter η represents the learning-paced parameter, which controls the learning method to select the appropriate samples to be present at the recent learning stage. As the training process proceeds, the parameter η is increased to incorporate the "easy-to-hard" idea into the training process, namely, the large value of η means that the easy samples are preferred to be selected.

3.2 Instance Segmentation Mask Initialization and Instance Cluster Generation

After viewpoint estimation, we adopt a co-segmentation approach [12] to initialize the segmentation masks of the instances. Due to large variations in viewpoints, shapes, textures, and sizes (see Fig. 1), it is difficult to implement such segmentation process on all instances from the same category. To this end, we build a refined two-stage clustering strategy based on the viewpoint and appearance (see Fig. 4). Our approach gradually decomposes the entire category-specific instance collection into multiple clusters with less

1. In Eq. (2), $\rho(\mathbf{p}_n, \mathbf{z}_n) = \rho(\mathbf{p}_n|\mathbf{z}_n)\rho(\mathbf{z}_n)$, where $\rho(\mathbf{p}_n|\mathbf{z}_n)$ is Gaussian (as given by Eq. (1)), and $\rho(\mathbf{z}_n) = \mathcal{N}(\mathbf{z}_n|0, \mathbf{I})$.

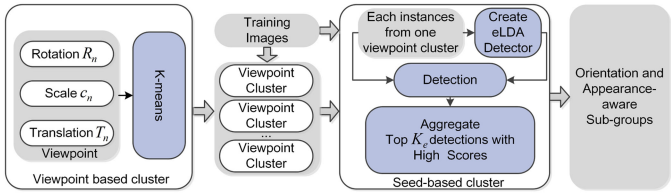


Fig. 4. Illustration of the two-stage clustering approach.

intra-group variations, which allows us to obtain robust priors from such subgroups and then use them to build the global shape information. Considering viewpoint variations, we use the camera parameters $\{c_n, \mathbf{R}_n, \mathbf{t}_n\}$ from Section 3.1 to describe each object instance and perform the K-means clustering method to separate the entire category-specific instance collection into δ viewpoint clusters, $\{\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, ..., \mathcal{G}^{(\delta)}\}$. Each cluster contains the instances from similar viewpoints. Considering shapes, textures, and sizes variations, we use the seed-based clustering approach [12] to obtain subgroups from each viewpoint cluster. This is due to the superior capability of the seed-based clustering approach in grouping visually coherent instances together. Specifically, we first use each instance within a certain viewpoint cluster as a seed and then build groups by detecting similar instances from the remaining instances. For implementation, we train the exemplar detectors eLDA [30] based on the HOG features of each instance, and then use each detector to group similar instances by selecting the top K_e detections with the highest detection scores.

For each training subgroup, the co-segmentation approach [12] can then be used to generate the initial segmentation masks of instances. Basically, the segmentation problem is formulated as a classical graphcut problem [31] that labels every pixel in the input images as the foreground or background. This graph cut problem is then solved by minimizing the energy function consisting of three terms: an image-level unary potential term, a cluster-level unary potential term, and a pairwise potential term. Specifically, the image-level unary potential term discovers the potential object regions by the appearance model specific to an image, the cluster-level unary potential term discovers the potential object regions by the appearance model shared between all images in the cluster, the pairwise potential term discovers the object boundaries between the object regions and image background.

3.3 3D Shape Reconstruction

This subsection presents the approach for learning the weakly supervised 3D object shape model based on a confidence weighting scheme (See Fig. 5).

3.3.1 Formulation

We use the following notations to present our 3D object reconstruction approach. Denote $\mathbf{O}$ as the 2D mask (i.e., silhouette) of the object in each image and $\mathbf{v} = [v_1^{(1)}, ..., v_{n_1}^{(1)}, v_1^{(2)}, ..., v_{n_2}^{(2)}, ..., v_{n_\delta}^{(\delta)}]$ as the confidence weights for all training subgroups, where n_c, $c = \{1, 2, \cdots, \delta\}$ indicates the number of subgroups in $\mathcal{G}^{(c)}$. Notice that in the previous works, e.g., [9], [10], the 2D object masks $\mathbf{O}$ are obtained by manual annotation. However, in this work, the manually annotated

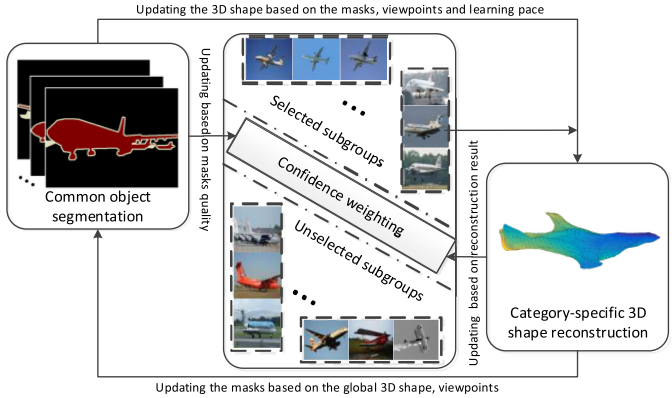


Fig. 5. Examples to illustrate the newly proposed confidence weighting-based 3D shape reconstruction scheme.

2D object masks $\mathbf{O}$ are not provided. Thus we need to infer the 2D object masks by introducing the common object segmentation process into the learning framework. $S = (\overline{\mathbf{Sh}}, \Gamma)$ is the 3D object shape model, where $\overline{\mathbf{Sh}}$ corresponds to the category-level 3D mean shape and $\Gamma = \{\gamma_1, \gamma_2, ..., \gamma_\omega\}$ denotes a set of deformation bases, ω is the number of the deformation bases. The 3D mean shape and the deformation bases are shared for all the images that contain instances belonging to the same object category.

For each training subgroup, we propose a soft-diverse confidence weighting variable that dynamically decides whether certain subgroup would be selected to participate in the learning process when constructing the 3D shape models. Specifically, the 3D shape reconstruction process is an iterative optimization process, which alternatively optimizes the 3D shape model based on the confident training subgroups selected from the previous iteration and updates the confidence weights of each training subgroup to select confident training subgroups for the next iteration. This learning procedure can well address the learning ambiguity issue due to noisy training instances, i.e., the instances with less-confident 2D segmentation masks. We formulate this new confidence weighting-based 3D shape reconstruction scheme as the following optimization problem:

$$\begin{aligned} &\min E_{GL} + E_{IA} + f(\mathbf{v}; \lambda, \vartheta), \\ s.t.\ \ &\mathbf{Sh}_n = \overline{\mathbf{Sh}} + \sum\nolimits_t \alpha_{n,t}\gamma_t, \end{aligned} \tag{3}$$

where $\alpha_{n,t}; t = \{1, ..., \omega\}$ denotes the deformation weight of the tth deformation base for the nth training instance. E_{GL} is a global energy term to regularize the learnt mean 3D shape and deformable bases, while E_{IA} is an instance-aware energy term to regularize the 3D shapes inferred on each training instance. $f(\mathbf{v}; \lambda, \vartheta)$ is the confidence weighting regularizer, which produces the soft confidence weights $\mathbf{v}$ for all training subgroups.

The Global Energy Term. Similar with [9], the global energy term E_{GL} contains a local consistency term E_{lc} and a deformation penalization term E_{pd}:

$$E_{GL} = E_{lc} + E_{pd}, \tag{4}$$

where

$$E_{lc}(\overline{\mathbf{Sh}},\Gamma)=\sum\nolimits_x\sum\nolimits_{y\in N(x)}((\parallel \overline{Sh}_x-\overline{Sh}_y\parallel-\varrho)^2 + \sum\nolimits_t\parallel \gamma_{t,x}-\gamma_{t,y}\parallel^2), \quad E_{pd}(\alpha,\Gamma)=\sum\nolimits_n\sum\nolimits_t\parallel \alpha_{n,t}\gamma_t\parallel_F^2. \tag{5}$$

In Eq. (5), the variable $\gamma_{t,x}$ is the xth point on the tth basis, and $\overline{Sh}_x$ is the three-dimensional coordinate of the xth point[2] on $\overline{\mathbf{Sh}}$. ϱ indicates the mean squared displacement between the neighborhoods $N(\cdot)$ of each point, which encourages all surfaces to have similar sizes. The neighborhood $N(\cdot)$ is defined based on the spatial distance. Different from the local consistency term E_{lc} which restricts arbitrary deformations, E_{pd} penalizes the L_2 norm of the deformation parameter α in order to prevent unreasonable large deformations.

The Instance-Aware Energy Term. The instance-aware energy term E_{IA} contains three terms, which are the appearance consistency term E_{sc}, the 3D shape normal smoothness term E_{ns}, and the common object segmentation term E_{COS}, respectively:

$$E_{IA}=\sum\nolimits_c\sum\nolimits_i\left(v_i^{(c)}\frac{1}{K_e}\sum\nolimits_{m=1}^{K_e}E_{sc}^{c,i,m}+E_{ns}^{c,i,m}+E_{COS}^{c,i,m}\right), \tag{6}$$

where c, i, and m indicate the indexes of the cluster, subgroup, and instance, respectively. Following [9], for each instance, we have:

$$E_{sc}(\mathbf{Sh};\mathbf{O},\pi)=\sum_{\Omega(q,\mathbf{O})>0}\triangle^1(q,\mathbf{O})+\sum_{p\in\mathbf{O}}\triangle^2(p,\pi(\mathbf{Sh})), \quad E_{ns}(\mathbf{Sh})=\sum\nolimits_x\sum\nolimits_{y\in N(x)}(1-\overrightarrow{\mathcal{N}}_x\cdot\overrightarrow{\mathcal{N}}_y), \tag{7}$$

where q denotes the point on the 2D projection of the 3D shape $\pi(\mathbf{Sh})$, $\Omega(q,\mathbf{O})$ refers to the Chamfer distance between the point q and the 2D object mask $\mathbf{O}$, $\triangle^1(q,\mathbf{O})$ indicates the squared distance of pixel q to its nearest neighbor in the set $\mathbf{O}$, $p\in\mathbf{O}$ denotes the point on the 2D object mask $\mathbf{O}$, $\triangle^2(p,\pi(\mathbf{Sh}))$ indicates the squared average distance of pixel p to its two nearest neighbors in the 2D projection $\pi(\mathbf{Sh})$ of its shape $\mathbf{Sh}$. Here $\pi(\mathbf{Sh})$ is defined as:

$$\pi(\mathbf{Sh})=c\,\mathbf{R}\,\mathbf{Sh}+\mathbf{t}\,\mathbf{1}^T, \tag{8}$$

where $\{c,\mathbf{R},\mathbf{t}\}$ are the camera viewpoint parameters obtained from Section 3.1.

According to [9], the first term in E_{sc} penalizes the 3D shape points which are outside the corresponding 2D object mask after projection, while the second term in E_{sc} encourages the points on the 2D object mask to pull nearby projected points towards them. Different from E_{sc}, the normal smoothness term E_{ns} places a cost on the variation of normal directions in a local shape neighborhood to ensure the shape change tends to be locally smooth (see Eq. (7)). In Eq. (7), $\overrightarrow{\mathcal{N}}_x$ denotes the normal for the xth point in $\mathbf{Sh}$. It is computed by fitting planes to the local point neighborhoods. The common object segmentation term E_{COS} is defined in Eq. (12), which is treated as a constant in Eq. (6).

The Confidence Weighting Regularizer. The confidence weighting regularizer term $f(\mathbf{v};\lambda,\vartheta)$ produces the soft confidence weights $\mathbf{v}$ for all training subgroups, which is defined as:

$$f(\mathbf{v};\lambda,\vartheta)=-\lambda\sum\nolimits_c\sum\nolimits_i v_i^{(c)}-\vartheta\sum\nolimits_c\sqrt{\sum\nolimits_i v_i^{(c)}}, \tag{9}$$

where the first term (referred to as the sample easiness term) tends to select samples (or training subgroups) with smaller loss values [2] while the second term (referred to as the sample diversity term) favors selecting training subgroups from different viewpoint clusters [2]. λ and ϑ are two parameters imposed on these two terms. With this confidence weighting regularizer, our approach tends to assign high confidence weights to the training subgroups with smaller learning loss and from diverse viewpoint clusters, which is critical for learning good 3D object shape models under the weak supervision. Without the variable $\mathbf{v}$ and this regularizer term, Eq. (3) degenerates to the objective function in [9], which learns the 3D object models from fully annotated 2D images.

3.3.2 Optimization

Given the training subgroups $\{G_i^{(c)};i=1,...,n_c,c=1,...,\delta\}$, the 3D shape models $S=(\overline{\mathbf{Sh}},\Gamma)$ are inferred by optimizing the energy function in Eq. (3). Similar with [2], [32], [33], the solution of Eq. (3) can be approximately obtained by alternatingly optimizing the confidence weights $\mathbf{v}$ and the shape models $(\overline{\mathbf{Sh}},\Gamma)$.

Optimize $(\overline{\mathbf{Sh}},\Gamma)$ *with Fixed* $\mathbf{v}$. The goal of this step is to update the category-specific 3D shape model. To relax the optimization process, we treat E_{COS} as a constant value that has been obtained from the previous segmentation phase. Then, the energy function in Eq. (3) degenerates to the following form:

$$\min_{\overline{\mathbf{Sh}},\Gamma,\boldsymbol{\alpha}} E_{lc}+E_{pd}+\sum\nolimits_c\sum\nolimits_i\sum\nolimits_m\left(E_{sc}^{c,i,m}+E_{ns}^{c,i,m}\right), \quad s.t.\ \ \mathbf{Sh}_n=\overline{\mathbf{Sh}}+\sum\nolimits_t\alpha_{n,t}\gamma_t. \tag{10}$$

To optimize this highly non-convex and non-smooth object function, we follow [9] to infer the optimal mean shape and the deformation basis by using block coordinate descent on $\overline{\mathbf{Sh}}$, Γ and α, in which sub-gradient is computed over the training samples. For initialization, we use the mean shape with a soft visual hull, which is computed by using the selected training subgroups. We initialize the deformation bases and deformation weights randomly.

Optimize $\mathbf{v}$ *with fixed* $(\overline{\mathbf{Sh}},\Gamma)$. This step updates the weights imposed on all training subgroups to reflect their confidence for learning of the 3D shape model in the next iteration. In this case, the energy function Eq. (3) is reformulated as follows:

$$\min_{\mathbf{v}} E_{IA}+f(\mathbf{v};\lambda,\vartheta), \quad s.t.\ \ \mathbf{Sh}_n=\overline{\mathbf{Sh}}+\sum\nolimits_t\alpha_{n,t}\gamma_t. \tag{11}$$

According to [2], Eq. (11) is convex as both the L1-norm term (i.e., the first regularizer in Eq. (11)) and the anti-group sparsity term (i.e., the second regularizer) are convex. By

2. We denote x and y as the indexes for the points on each 3D shape, while p and q as the indexes for the points on each 2D image.

satisfying the KKT (Karush-Kuhn-Tucker) conditions of the Lagrangian, the explicit global optimum of Eq. (11) can be efficiently calculated. We provide the explicit solution of Eq. (11) in Algorithm 1 of the Supplementary Material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2019.2949562.

3.4 Common Object Segmentation Using 3D Shape Projection Prior

We use the category-specific 3D shape models to provide top-down priors for common object segmentation. Specifically, for each training subgroup containing K_e instance images, our goal is to obtain each object mask $\mathbf{O}$ by labeling each pixel to be foreground (i.e., $l_p = 1$) or background (i.e., $l_p = 0$), where p denotes the pixel location in each image. This labeling problem is solved by minimizing the following energy function over the pixels and labels:

$$E_{COS} = E_I + E_W + E_{TD}, \tag{12}$$

where the term E_I [12] is the unary potential from an appearance model that is specific to each instance:

$$E_I(p; A) = -\log \rho\big(l_p; \mathbf{x}_p, A\big), \tag{13}$$

where $\rho\big(l_p; \mathbf{x}_p, A\big)$ measures the likelihood of a pixel with its RGB color feature $\mathbf{x}_p$ taking the label l_p according to its appearance model A. Here A consists of two Gaussian mixture models (GMMs) over the RGB color space as defined in [12] i.e., one for the foreground (when $l_p = 1$) and another for the background (when $l_p = 0$). The appearance model is learned by using the pixels inside and outside the segmentation mask, which is inferred from the previous iteration.

The term E_W [12] is the pairwise potential defined as:

$$E_W\big(p, q; l_p, l_q\big) = \delta\big(l_p \neq l_q\big) e^{-\beta \left\| \mathbf{x}_p - \mathbf{x}_q \right\|^2}, \tag{14}$$

where $\delta(\cdot)$ is the logistic function. Eq. (14) penalizes two pixels (p and q) when they are assigned with different labels but having an intervening contour (IC) between them [34].

The term E_{TD} is the top-down prior term that enforces the obtained segmentation masks within each subgroup to be consistent. This is modeled as the top-down shape priors over pixels:

$$\begin{aligned} &E_{TD}(p; \mathbf{SM}, \mathbf{PM}) = \\ &-\log \rho(l_p|\mathbf{SM}, p) - \log \rho(l_p|\mathbf{PM}, p), \end{aligned} \tag{15}$$

where $\mathbf{SM}$ is the average segmentation mask of the instances in the subgroup. The term $-\log \rho(l_p|\mathbf{SM}, p)$ is the prior probability that each pixel belongs to the foreground or background, given the pixel location and $\mathbf{SM}$. Similarly, $\mathbf{PM}$ is the average projection shape mask of the instances in the subgroup, which is obtained by using the 3D shape model $S = \big(\overline{\mathbf{Sh}}, \Gamma, \alpha\big)$ from the previous iteration:

$$\begin{aligned} &\mathbf{PM} = \frac{1}{K_e} \sum\nolimits_m \big(c_m \mathbf{R}_m \mathbf{Sh}_m + \mathbf{t}_m \mathbf{1}^T\big), \\ &\mathbf{Sh}_m = \overline{\mathbf{Sh}} + \sum\nolimits_t \alpha_{m,t} \gamma_t, \end{aligned} \tag{16}$$

where the viewpoint parameters $\{c_m, \mathbf{R}_m, \mathbf{t}_m\}$ are obtained from Section 3.1, the 3D shape model $\big(\overline{\mathbf{Sh}}, \Gamma\big)$ and the corresponding deformation parameters $\{\alpha_{m,t}\}$ are obtained from Section 3.3. The term $-\log \rho(l_p|\mathbf{PM}, p)$ effectively introduces the top-down prior provided by the category-specific 3D shape models for common object segmentation. After the 3D object reconstruction process, we use the learned 3D shape model to obtain the average projection shape mask $\mathbf{PM}$ and treat it as a constant in Eq. (12). Then, the overall energy in Eq. (12) can be conveniently minimized by using the graph-cut algorithm [31].

4 Experimental Evaluation

4.1 Tasks and Evaluation Metrics

We conduct individual experiments for three main components: the category-specific 3D shape models, the common object segmentation masks, and the viewpoint estimation.

To evaluate category-specific 3D shape models, we use the following procedure. After the learning procedure with the newly proposed algorithm, we follow the testing process in [9] to use the learned 3D shape models, which contains a 3D mean shape and a set of deformation bases, to infer the deformation parameters and fit the 2D silhouette of the objects in each testing image. Then we evaluate the quality of the reconstructed 3D object shapes, which are obtained by combing our learnt 3D mean shape and deformation bases with the deformation parameters inferred on each testing image, based on two metrics, i.e., the Mesh Error [35] and the Depth Error [9].[3]

For evaluating the generated segmentation masks, we adopt the standard intersection-over-union (IOU) metric to compare each segmentation mask and the corresponding ground-truth mask on the training set.

For viewpoint estimation, we use geodesic distance to compute the difference between our estimated viewpoints and the 3D annotated viewpoints from the PASCAL3D+ dataset [36]. Then, we employ two complementary metrics to evaluate the viewpoint estimation accuracy. The first one is the median error, which refers to the median value of all the geodesic distances between the predictions and the ground-truth rotation matrices. The second one is the accuracy at θ, which denotes the proportion of instances whose predicted viewpoints are within a fixed range (i.e., θ degree) of the corresponding ground-truth viewpoints. We also follow [9] to denote the threshold $\theta = \frac{\pi}{6}$.

More detailed descriptions of the above mentioned evaluation metrics can be referred to the Supplementary Material, available online.

4.2 Dataset

We use 10 rigid object categories selected from the challenging PASCAL VOC 2012 benchmark [37], which is widely used for object detection, semantic segmentation and many another computer vision tasks. However, the label from the PASCAL VOC dataset does not include the annotation of object viewpoints and 3D shapes. In order to evaluate the performance of different methods used for 3D shape reconstruction and

3. As this work focuses on the learning process of the 3D object shape models, following the testing process of the previous works in the aforementioned way could help present a more clear comparison evaluation.

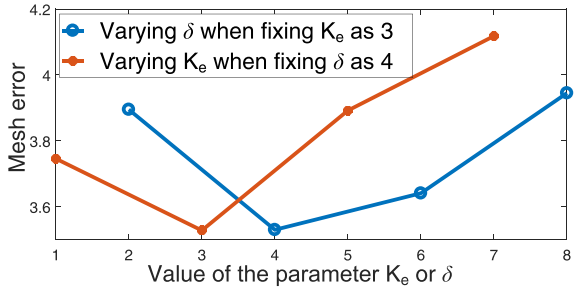


Fig. 6. Analysis of the performance (mesh error) variation when using different parameters δ and K_e in the proposed learning framework.

viewpoint estimation, we adopt the 3D CAD models and the viewpoint annotation, which are provided in the PASCAL3D + dataset [36]. To the best of our knowledge, the PASCAL 3D + dataset is the most comprehensive dataset as it contains a large number of images with 3D shape and viewpoint annotation. In our experiments, we use the publicly available keypoints during the learning phase and adopt the ground-truth segmentation masks for evaluation. During training, we only use the images that contains one object.

4.3 Parameters Setting and Initialization

To implement the proposed robust NRSfM method, we set the number of deformable base $D = 5$ and other parameters (i.e., c_n, $\mathbf{R}_n$, $\mathbf{t}_n$, $\hat{\mathbf{U}}$ and σ^2) are randomly-initialized as in [15]. For initialization in our proposed robust NRSfM, we need to determine the initial parameter for each sample. To this end, we first perform NRSfM for 10 iterations by using randomly initialized parameters. Then select 80 percent of samples for initial training, which will be then increased by 10 percent after each iteration. Please refer to the Supplementary Material, available online for the performance variation using different hyperparameters.

According to the experimental analysis shown in Fig. 6, we set the number of viewpoint clusters δ as 4 and the number of top detections K_e as 3. From Fig. 6, we observe that the final performance cannot be always improved by increasing the number of viewpoint-specific clusters δ. Specifically, although the viewpoint consistency within each cluster would increase when using more clusters, the appearance consistency within each cluster might be hurt, which will also influence the subsequent appearance-based clustering process.

For 3D object reconstruction, we set the number of deformable bases as $\omega = 5$. The mask of each object and the 3D mean shape for each category are initialized by using the common segmentation results in [12] and the visual hulls method in [9] respectively. During the learning process, we propose a five-level approach, which starts from a smaller confidence parameter to infer a coarse category-level mean 3D shape model and then gradually increases the confidence parameter to use more complex subgroups for enriching the details of the 3D shape. The confidence parameters λ and ϑ are jointly determined by the number of subgroups involved in the training process and the value of the energy function $\frac{1}{K_e}\sum_{m=1}^{K_e} E_{sc}^{c,i,m} + E_{ns}^{c,i,m} + E_{COS}^{c,i,m}$. During the learning process, we select 50 percent subgroups in the first iteration and then gradually increase the amount of selected subgroups by 10 percent per iteration. After each iteration, we first sort the training subgroups according to the values of their energy function $\frac{1}{K_e}\sum_{m=1}^{K_e} E_{sc}^{c,i,m} + E_{ns}^{c,i,m} + E_{COS}^{c,i,m}$ in ascending order and choose the top ranked subgroups. Then, we set λ as the maximum loss value of the selected subgroups and ϑ as 20 percent of standard deviation for the loss values of all training subgroups. Please refer to the Supplementary Material, available online for the performance variation using different hyperparameters.

The appearance model A_m and the RGB color feature $\mathbf{x}_{m,p}$ are obtained by the Chen et al. in work [12]. In the first iteration, the foreground-background prior $\mathbf{SM}$ is provided by the bounding box and the top-down prior term E_{TD}, which only considers the prior probability term $-\log \rho(l_p|\mathbf{SM}, p)$. In the subsequent iterations, once we finish 3D object reconstruction process, the masks in each training subgroup are updated by using the planar projection of the learnt 3D shape.

4.4 Viewpoint Estimation

We evaluate the proposed viewpoint estimation method by comprehensively comparing it with the state-for-the-art structure-from-motion method EM-PPCA [15], PF-NRSfM [39] and the Rigid-SFM method [41]. The quantitative comparison results on the benchmark datasets are shown in Table 1, from which we observe that the proposed learning strategy can improve the viewpoint estimation performance in terms of the median error and the accuracy at $\Pi/6$. In Table 1, we additionally compare our method with a baseline "EM-PPCA-CH". The "EM-PPCA-CH" baseline uses a mask constraint (i.e., the convex hull based on keypoints) in

TABLE 1
Comparison Between the Estimated Viewpoints From Our Approach and the Two Baseline Methods Vanilla EM-PPCA and Rigid-SFM in Terms of Median Error (MedErr) and Accuracy at $\Pi/6$ (Acc $\Pi/6$)

	Categories	aero	bike	boat	bus	car	chair	mbike	sofa	train	tv	mean
MedErr	PF-NRSfM [39]	18.42	19.35	86.29	13.52	19.18	11.14	27.60	29.06	29.82	18.09	27.25
	Rigid-SFM [41]	17.14	18.07	76.95	13.47	15.92	9.19	25.29	28.02	29.58	17.28	25.09
	EM-PPCA [15]	16.25	16.07	67.83	12.82	10.85	8.99	23.37	21.10	22.70	15.41	21.54
	EM-PPCA-CH	15.96	34.19	67.83	12.89	10.79	8.99	**18.46**	21.58	23.04	**13.24**	22.7
	Ours	**15.90**	**14.99**	**62.82**	**10.89**	**7.21**	**8.15**	19.20	**17.36**	**20.58**	13.64	**19.07**
Acc $\Pi/6$	PF-NRSfM [39]	0.54	0.76	0.02	0.80	0.76	0.85	0.57	0.53	0.50	0.74	0.61
	Rigid-SFM [41]	0.55	0.79	0.03	0.80	0.79	0.89	0.60	0.56	0.52	0.77	0.63
	EM-PPCA [15]	0.57	0.82	0.10	0.85	0.97	0.89	0.63	0.69	0.63	0.81	0.70
	EM-PPCA-CH	**0.58**	0.38	**0.12**	0.86	0.97	0.89	**0.76**	**0.76**	0.63	**0.86**	0.68
	Ours	**0.58**	**0.83**	**0.12**	**0.88**	**0.98**	**0.92**	0.75	**0.76**	**0.67**	0.82	**0.73**

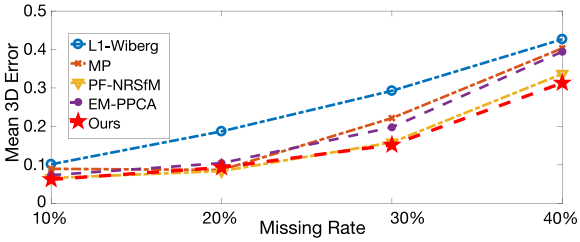


Fig. 7. Reconstruction performance comparison on missing data by using the proposed method, one baseline method and four state-of-the-art methods MP [38], EM-PPCA [15], PF-NRSfM [39], and L1-Wiberg [40]. Notice that the missing data rates are 10, 20, 30, and 40 percent, respectively.

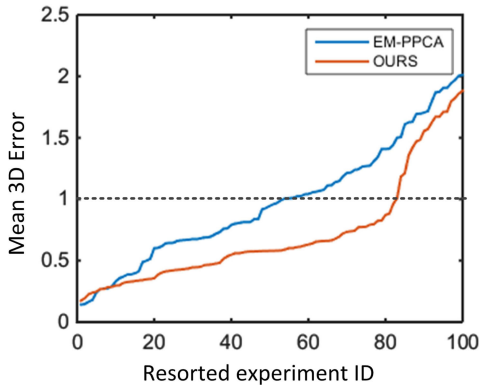


Fig. 8. Statistic of the EM-PPCA method over 100 times experiments and our proposed method with random initialization. The results with larger than 1 Mean 3D Error are considered as bad local minimums.

its objective function, while our method does not use this constraint. Therefore, for the categories like "airplane", "boat", "motorbike" and "tv", this baseline can achieve slightly better performance than ours. However, for the "bike" category, the objects have complex shapes and structure. Thus, the convex hulls formed by keypoints can hardly reflect the real structure of objects. In this case, this method achieves much worse performance than the EM-PPCA baseline as well as our approach.

To demonstrate the robustness of the proposed view point estimation method for handling the missing data, we compare our proposed robust NRSfM method with 4 state-of-the-art methods, including three structure-from-motion methods (i.e., MP [38], EM-PPCA [15] and PF-NRSfM [39]) and a robust matrix approximation with missing data method (i.e., L1-Wiberg [40]), and a standard NRSfM baseline [9]. We perform this experiment on the simulated missing data (a 900 by 20 matrix), which is generated by the data generation method in [15], in which the missing data rates are 10, 20, 30 and 40 percent, respectively. Following the previous work [39], we use the normalized mean 3D errors to evaluate the performance of each method. The quantitative comparison results are shown in Fig. 7. Compared with the state-of-the-art methods and the baseline, our robust NRSfM method can achieve competitive performance and is also robust to missing data. While the PF-NRSfM method [39] also achieves good performance on the synthetic data, it cannot achieve as good performance as our approach in more challenging scenarios like the Pascal dataset (see Table 1s).

To further analyze the capacity of the proposed robust NRSfM method for avoiding arriving at a bad local minimum solution, we compare our approach with EM-PPCA [15] by performing each method for 100 times with random initialization. The experimental results are shown in Fig. 8, where we observe that our method can achieve better average performance than EM-PPCA. In addition, over all these 100 experiments, EM-PPCA arrives at a bad local minimum for 47 times, where a bad local minimum is defined as the one that has larger error than 1 Mean 3D Error [39]. In contrast, our method arrives at a bad local minimum for 18 times only. The experiments demonstrate the effectiveness of our robust RNSfM method for avoiding arrives at a bad local minimum solution.

4.5 Results for 3D Shape Reconstruction

In this section, we first compare the 3D Shape Reconstruction results of our approach with the baseline methods: ("Ours-nSV-nSR-nCS", "Ours-nSR-nCS" and "Ours-nCS"). Specifically, "Ours-nSV-nSR-nCS" directly utilizes the initial co-segmentation masks of all the training images to reconstruct the category-specific 3D shape models without using our proposed viewpoint estimation and cross task cooperation methods. "Ours-nSR-nCS" reconstructs the 3D models with the initial co-segmentation masks and without using cross task cooperation for reconstruction. "Ours-nSR" jointly implements category-specific 3D shape reconstruction and common object segmentation but without using our newly proposed confidence weighting scheme. Furthermore, we also compare the performance by using different confidence weighting regularizers. Specifically, we build the baseline by replacing the $l_{0.5,1}$ norm-based soft-diverse confidence weighting regularizer in Eq. (9) with the l_1 norm-based regularizer used in [28], which only produces binary information of the important weight v. We name this baseline as "Hard SPL". The quantitative evaluation results of these baseline methods are reported in Table 2.

From Table 2, we have the following observations: 1) The confidence weighting scheme under the self-paced learning framework provides an effective way to improve the weakly supervised learning process for 3D object shapes (see the comparison between "Ours", "Hard SPL" and "Ours-nSR"). 2) Our newly proposed viewpoint estimation method can improve the reconstruction performance (see the comparison between "Ours-nSV-nSR-nCS" and "Ours-nSR-nCS"). 3) The joint category-specific 3D shape and common object segmentation method can help 3D object reconstruction and improve the learning performance (see the comparison between "Ours-nSR-nCS" and "Ours-nSR"). 4) By additionally considering the softness and diversity criterion for the weighting regularizer, The approach using the $l_{0.5,1}$ norm-based self-paced regularizer further outperforms the approach using alternative conventional l_1 norm-based regularizer (see the comparison between "Hard SPL" and "Ours"). 5) The proposed approach ("Ours") achieves best overall performance and the outperforms other baseline methods on all these categories in terms of the mesh error and depth error.

In addition, we compare the proposed approach with a baseline method "KP-CH". Instead of using the ground-truth masks, the "KP-CH" method obtains the 3D shape model from the convex hulls of keypoints by using the method in [9]. The reconstruction results are reported in Table 2, and some examples are shown in Fig. 9. We observe that the baseline method "KP-CH" cannot achieve acceptable performance

TABLE 2
Comparison of the Learnt 3D Shape Models Between the Proposed Approach and These From the Weakly Supervised Baseline Methods

	Categories	aero	bike	boat	bus	car	chair	mbike	sofa	train	tv	mean
MeshErr	Coarse-Seg	2.36	4.30	6.21	4.12	4.36	4.04	3.12	8.56	12.11	10.87	6.00
	Ours-nSV-nSR-nCS	2.04	4.09	4.29	3.21	2.34	3.36	2.34	6.36	8.83	9.49	4.64
	Ours-nSR-nCS	1.97	3.37	4.05	3.10	2.33	3.01	2.28	6.33	8.67	9.16	4.43
	Ours-nSR	1.89	3.28	3.96	2.24	2.32	2.55	2.25	6.32	8.37	7.73	4.09
	KP-CH	2.62	4.24	5.98	4.25	3.24	4.08	3.14	8.29	10.34	9.52	5.37
	Hard SPL	1.74	2.37	3.64	2.23	2.31	2.46	2.22	6.01	8.30	3.89	3.52
	Ours	**1.73**	**2.29**	**3.56**	**2.14**	**2.12**	**2.28**	**2.08**	**5.72**	**7.78**	**3.74**	**3.32**
DepthErr	Coarse-Seg	12.65	19.22	22.39	23.38	18.41	16.40	14.24	35.77	43.29	38.11	24.38
	Ours-nSV-nSR-nCS	10.77	14.73	17.13	18.51	11.22	10.72	11.73	26.60	37.50	36.84	19.58
	Ours-nSR-nCS	10.74	14.35	17.09	18.23	11.17	10.50	11.41	26.32	37.49	33.98	19.13
	Ours-nSR	10.63	13.18	17.01	18.04	11.06	10.40	11.39	26.30	37.42	21.03	17.65
	KP-CH	13.45	18.93	21.23	24.85	13.34	17.71	15.28	34.76	38.10	36.96	23.46
	Hard SPL	10.43	10.34	16.64	18.01	10.98	10.35	10.99	26.02	37.39	14.52	16.57
	Ours	**10.42**	**10.23**	**15.59**	**16.55**	**10.59**	**10.12**	**10.77**	**25.65**	**35.51**	**13.67**	**15.91**

Detailed description of the compared methods can be referred to as in Section 4.5.

due to lack of complex details in convex hulls, which also demonstrates the importance of our fine-segmented masks for reconstructing fine 3D shapes.

We also compare the 3D shape models in our newly proposed approaches with those from several state-of-the-art methods, including Tulsiani et al. [9], Vicente et al. [13], Twarog et al. [42], Barron et al. [43]. When compared with our approach, all these state-of-the-art methods need to additionally use a large amount of manually labeled segmentation masks. However, from the experimental results in Table 3, we observe that our method achieves encouraging performance. In the first seven categories in Table 3, our approach is comparable with these state-of-the-art methods, which utilize stronger supervision (i.e., the ground-truth-masks). We also note that our newly proposed method cannot perform well on the categories with complex scenes (e.g., "train" and "sofa") due to lack of enough instances. We show some experimental results in Fig. 10, which include the successful and failure cases. As shown in Fig. 10, it is challenging for our approach to recover objects which have highly complex structure and contain inconsistent topology with the mean shape as it is hard to segment these objects and reconstruct their 3D shapes. It is the common challenge to handle such case in the field of 3D shape reconstruction even for the state-of-the-art methods with stronger supervision.

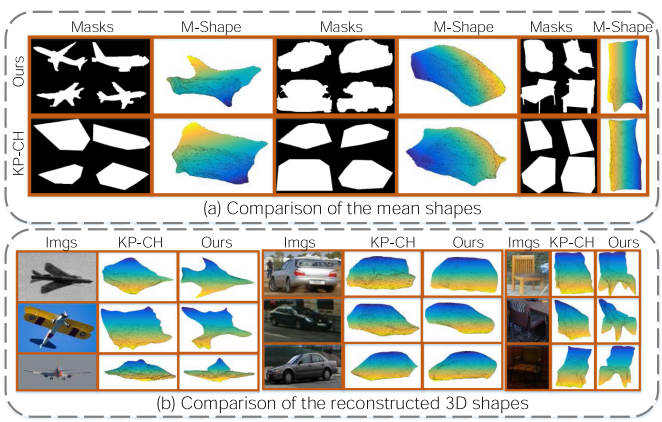


Fig. 9. Visual comparisons of the proposed 3D shape reconstruction method and the keypoints based 3D shape reconstruction method using convex hulls. The subfigure (a) shows the comparison of the mean shapes from our inferred segmentation masks and the mean shapes obtained based on the convex hull of the key points. Notice that in the "masks" columns of the subfigure (a) we only show four random examples of the segmentation masks from the corresponding category. The subfigure (b) shows the comparison of the 3D shapes reconstructed based on the convex hull of the key points and our final segmentation masks for each peculiar test instance.

Moreover, we compare our method with the Warpnet method [44], which is a recent point-matching based method that requires category labels for weakly-supervised learning of 3D shape model. For comparison, we implement the Warpnet method for one category (i.e., "car") in the PASCAL VOC dataset and build the pose-graph and match the image pairs from this category for 3D points reconstruction. Since Warpnet [44] reconstructs the 3D shape (in the form of 3D points) of the camera perspective for each object, we rotate the ground-truth shape into the camera view by using the ground truth viewpoints and evaluate the predicted shapes in terms of mesh error. When compared with Warpnet [44], our method achieves better performance on the car category, i.e., 2.12 (Ours) versus 2.21 (Warpnet [44]) in terms of mesh error. We also show the predicted 3D points in Fig. 11. For our method, we show the vertices of the reconstructed 3D mesh shapes. We observe our approach obtains better 3D shapes. There are two possible reasons: 1) We learn a parametric shape model for objects from each category. Such model is able to capture the intra-category shape variations [9]. 2) We update the 2D object masks during the learning process whereas the work in [44] fixes the imprecise object masks during their whole learning process.

We further analyze the efficacy of joint segmentation and reconstruction by comparing the performance of the proposed approach with a set of baseline methods, which include "Auto-Seg" and "Ours-nUS". The "Auto-Seg" baseline jointly performs the joint co-segmentation and 3D shape reconstruction by using [6] and [13], while "Ours-nUS" performs the naive co-segmentation method and the 3D shape reconstruction of our proposed framework without updating segmentation in each iteration. The quantitative comparison results are shown in Fig. 12, from which we can observe that: 1) The Auot-Seg baseline which directly uses the co-segmentation

TABLE 3
Comparison Between the Learned 3D Shape Models Obtained From the Proposed Approach and the State-of-the-Art (STA) Methods

	Categories	aero	bike	boat	bus	car	chair	mbike	sofa	train	tv	mean
MeshErr	Tulsiani et al. [9]	**1.72**	**1.78**	3.01	**1.90**	1.77	**2.18**	1.88	**2.13**	**2.39**	**3.28**	**2.20**
	Vicente et al. [13]	1.87	1.87	**2.51**	2.36	**1.41**	2.42	**1.82**	2.31	3.10	3.39	2.31
	Twarog et al. [42]	3.30	2.52	2.90	3.32	2.82	3.09	2.58	2.53	3.92	3.31	3.03
	Ours	1.73	2.29	3.56	2.14	2.12	2.28	2.08	5.72	7.78	3.74	3.32
DepthErr	Tulsiani et al. [9]	**9.51**	**9.27**	17.20	**12.71**	9.94	**7.78**	9.61	**13.70**	31.58	8.78	**13.01**
	Vicente et al. [13]	10.05	9.28	**15.06**	18.51	**8.14**	7.98	**9.38**	13.71	**31.25**	**8.33**	13.17
	Barron et al. [43]	13.52	13.79	20.78	29.93	22.48	18.59	16.80	18.28	40.56	20.18	21.49
	Ours	10.42	10.23	15.59	16.55	10.59	10.12	10.77	25.65	35.51	13.67	15.91

Notice that all the STAs require stronger supervision (i.e., the manually annotated segmentation masks of object instances) than the proposed approach.

masks to reconstruct 3D shapes cannot obtain satisfactory results, while our proposed strategy that simultaneously performs reconstruction and segmentation can effectively improve the 3D shape reconstruction performance. 2) the 3D object reconstruction results can be improved under our weakly supervised learning setting by gradually updating the segmentation masks along the learning iterations.

Finally, we qualitatively analyze the performance of 3D shape reconstruction by using coarse segmentations ("Coarse-Seg"). For "Coarse-Seg", we first train the 3D shape models by using the coarse segmentation results obtained from [45] and then we follow the 3D shape reconstruction method in [9] to use the learnt 3D shape models to fit the testing images. The results of the baseline method "Coarse-Seg" are reported in Table 3. When compared with this baseline, our method achieves better performance. Some examples of the predicted shapes by using the baseline method "Coarse-Seg" are shown in Fig. 13. From Fig. 13, we observe that our proposed method can recover satisfactory 3D shapes with fine details and complete appearances, whereas the recovered results are much worse by using the coarse segmentation method [45].

4.6 Results for Common Object Segmentation

First, we compare our proposed learning framework with three baseline methods "Hard SPL", "Ours-nSR-nCS" and "Ours-nSR" in terms of the IOU score. The experimental results are reported in the top part of Table 4, from which we

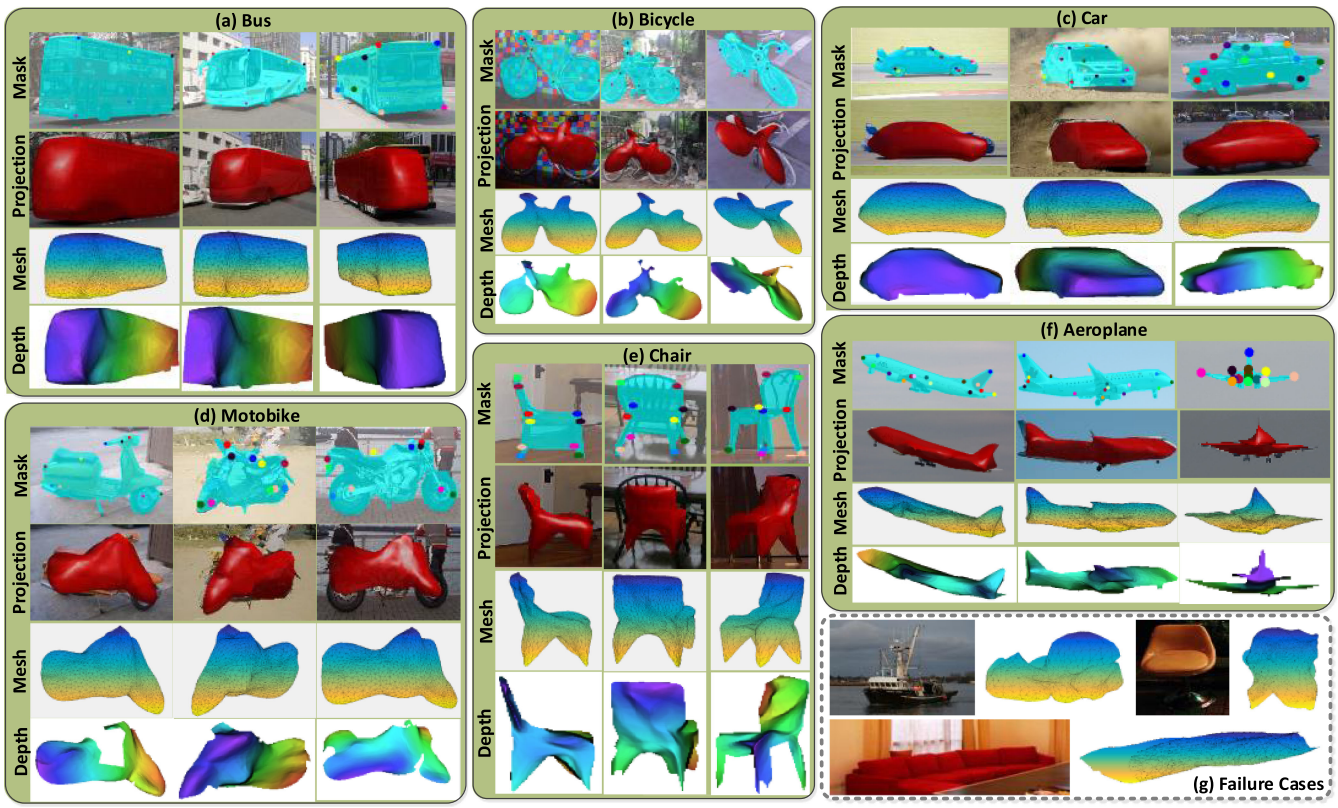


Fig. 10. Examples of the mesh maps, the depth maps, the manually annotated keypoints on different masks and the 3D shapes projected on the images by using our method. In subfigures (a)-(f), "Mask" shows the ground-truth mask together with the manually annotated keypoints, "Projection" shows the projection obtained by projecting the 3D shapes onto the images according to the estimated viewpoints (see Section 3.1), "Mesh" and "Depth" are the 3D shape mesh and the depth map according to its viewpoint, respectively. The subfigure (g) shows some failure cases, which are also challenging for the STA methods with stronger supervision.

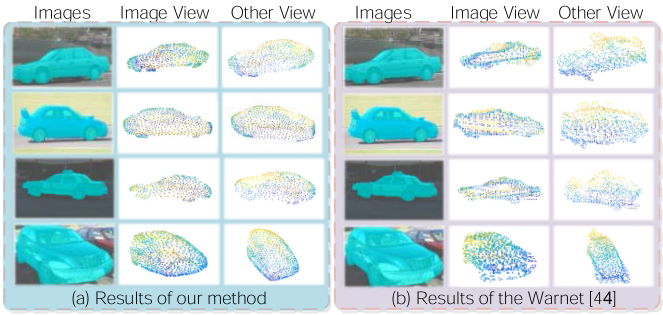


Fig. 11. Examples show the reconstruction results of our method and the Warpnet method [44], where the subfigure (a) shows the 3D points reconstructed by our method and the subfigure (b) shows the 3D points reconstructed by the Warpnet method. The reconstructed points are shown from the image view and the 45 degree elevation of the image view (i.e., other View).

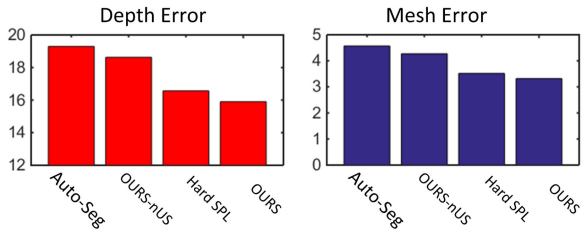


Fig. 12. Comparison of our proposed 3D shape reconstruction method and the automatic segmentation based 3D shape reconstruction method.

observe that: 1) it is beneficial to use the proposed viewpoint estimation and the confidence weighting-based cross task cooperation method for co-segmentation. 2) Basically, "Ours-nSR" outperforms "Ours-nSR-nCS" by jointly inferring the category-specific 3D shapes and common object segmentation masks, which is consistent with the expressiveness of the learned 3D models. 3) By additionally considering the softness and diversity criterion when designing the regularizer, the soft-diverse confidence weighting method "Ours" further improves the co-segmentation performance.

We also compare the segmentation masks for the obtained 3D shape models to segment the common objects via Eq. (13) with those obtained from three state-of-the-art object co-segmentation methods [6], [12], [45]. For fair comparison, we implement the state-of-the-art methods by using them to segment the instance images that are cropped according to the keypoint location. As shown in the bottom part of Table 4, the proposed approach achieves better performance than

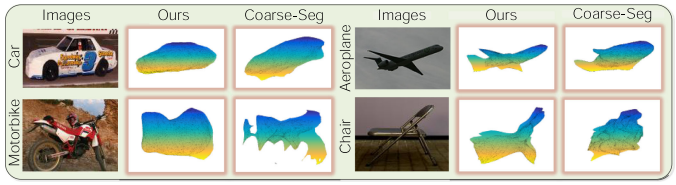


Fig. 13. Visual comparison of the 3D shapes obtained by our method (denoted as "Ours") and the coarse segmentation method [45] (denoted as "Coarse-Seg").

all the baseline and state-of-the-art methods in terms of the IOU score.

During the experiment, we observe that the segmentation results might benefit from using the foreground and background priors from the annotated keypoints. Considering that such annotated keypoints are actually used in our method, we further compare our method with the state-of-the-art co-segmentation methods by employing the convex hulls of keypoints at their initialization stages for more fair comparison, which forms the baseline methods "Chen-KP", "Joulin-KP" and "Quan-KP". We also report the IOU scores of the foreground-background-prior "BB" baseline, which uses the prediction of all the pixels in the GT bounding box for segmentation. This can be considered as the lower-bound of the segmentation results. Besides, we also report the IOU scores of the "NS" baseline, which employs the supervised segmentation method [46] to generate the segmentation masks. Thus, the 'NS" baseline can be considered as the upper-bound of the segmentation results to some extent. The experimental results are reported in Fig. 14, from which we observe that our approach can achieve better average performance when compared with STAs which takes advantage of keypoint annotation. It is worth mentioning that the performance improvement is not due to the keypoints but the 3D shape prior.

4.7 Analysis of Convergence and Time Consumption

First, we implement additional experiments to analyze the convergence of our approach. Specifically, we evaluate our method at each training stage. The performance is evaluated based on the value of the 3D object reconstruction objective function on the training data as well as the Mesh error and Depth error on the test data. The experimental results are shown in Fig. 15. From the left subfigure in Fig. 15, i.e., the Objective Energy Value versus Training iteration figure,

TABLE 4
Comparison Between the Segmentation Masks From Our Approach and Other Baselines and STAs in Terms of IOU

	Categories	aero	bike	boat	bus	car	chair	mbike	sofa	train	tv	mean
Baselines	Ours-nSR-nCS	0.755	0.590	0.672	0.822	0.791	0.685	0.725	0.862	0.643	0.671	0.722
	Ours-nSR	0.769	0.591	0.678	0.825	0.791	0.695	0.729	0.867	0.665	0.708	0.732
	Hard SPL	0.779	0.612	0.695	0.829	0.798	0.725	0.735	0.882	0.668	0.785	0.752
	Ours	**0.791**	**0.665**	**0.705**	**0.845**	**0.813**	**0.754**	**0.756**	**0.893**	**0.690**	**0.813**	**0.772**
STAs	Quan et al. [6]	0.729	0.481	0.644	0.764	0.788	0.608	0.743	0.831	0.666	0.648	0.690
	Chen et al. [12]	0.684	0.544	0.585	0.739	0.749	0.650	0.654	0.891	0.670	0.723	0.689
	Joulin et al. [45]	0.279	0.336	0.239	0.378	0.319	0.236	0.334	0.435	0.363	0.260	0.318
	Ours	**0.791**	**0.665**	**0.705**	**0.845**	**0.813**	**0.754**	**0.756**	**0.893**	**0.690**	**0.813**	**0.772**

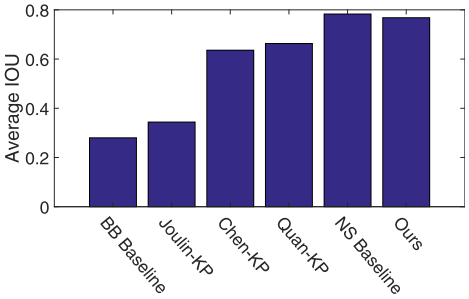


Fig. 14. Experimental comparison of the proposed method ("Ours") and three keypoint based STA co-segmentation methods (Joulin-KP, Chen-KP, and Quan-KP). Notice that the results of "Chen-KP", "Joulin-KP," and "Quan-KP" are initialized by connecting keypoints. The "NS-Baseline" is a supervised segmentation method [46].

we can observe that our approach tends to converge within five iterations. While the right two subfigures in Fig. 15 show that our approach could indeed obtain good performance by the model obtained from the fifth training iteration. These two subfigures do not show an obvious convergence property as in the left figure as there are variations between the training data and testing data.

Furthermore, we also analyze the time consumption by comparing our approaches with the most recent state-of-the-art method [9]. The experiments are performed 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 [9] (4.39 hours) as we need to additionally infer the segmentation masks. During testing, our method takes 38s per image, which is the same as [9]. In terms of time complexity, our method is worse at the training stage. However, our method still has advantages after considering other features including annotation costs. According to our statistics in [14], the average time cost for manually annotating image labels, keypoints, and segmentation masks are 1.2s, 4.4s, and 256.1s per image, respectively. The segmentation label takes 97.9 percent of the entire human effort for annotating 2D images. This indicates that our approach can save significant amount of human effort for 3D shape reconstruction.

5 Conclusion

In this paper, we have proposed a novel framework to jointly perform 3D object reconstruction and common segmentation based on weakly annotated images. Our confidence weighting strategy effectively guides and improves the training process of the proposed framework. Comprehensive experiments on PASCAL VOC have demonstrated the notable performance of the proposed framework. Our approach significantly reduces the manual annotation costs, and could make it unprecedented cheap for learning 3D shape models and thus potentially facilitate large-scale applications in the future.

Essentially, there are still several open problems, such as, lack of the keypoint labels [44], non-rigid objects, 3D object reconstruction in fine scale. We plan to address these issues by investigating an unsupervised framework for jointly estimating the viewpoints and poses, and perform fine scale 3D shape reconstruction based on the coarse scale 3D models, viewpoints and poses.

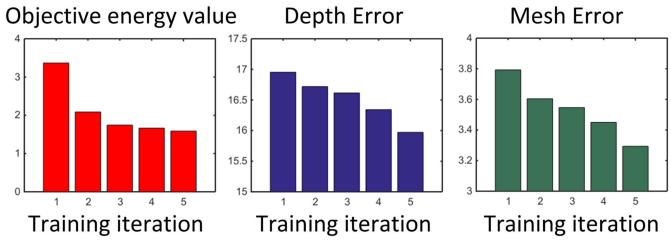


Fig. 15. Convergence analysis of our iterative learning procedure.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2017YFB1002201, the National Science Foundation of China under Grant 61876140, and the China Postdoctoral Support Scheme for Innovative Talents under Grant BX20180236.

References

[1] D. Zhang, J. Han, C. Li, J. Wang, and X. Li, "Detection of co-salient objects by looking deep and wide," *Int. J. Comput. Vis.*, vol. 120, no. 2, pp. 215–232, 2016.

[2] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 865–878, May 2017.

[3] G. Cheng, P. Zhou, and J. Han, "RIFD-CNN: Rotation-invariant and fisher discriminative convolutional neural networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2884–2893.

[4] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 379–387.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, vol. 1, 2015, pp. 91–99.

[6] R. Quan, J. Han, D. Zhang, and F. Nie, "Object co-segmentation via graph optimized-flexible manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 687–695.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

[9] S. Tulsiani, A. Kar, J. Carreira, and J. Malik, "Learning category-specific deformable 3D models for object reconstruction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 719–731, Apr. 2017.

[10] A. Kar, S. Tulsiani, J. Carreira, and J. Malik, "Category-specific object reconstruction from a single image," in *Proc. Comput. Vis. Pattern Recognit.*, 2015, pp. 1966–1974.

[11] J. Carreira, S. Vicente, L. Agapito, and J. Batista, "Lifting object detection datasets into 3D," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 7, pp. 1342–1355, Jul. 2016.

[12] X. Chen, A. Shrivastava, and A. Gupta, "Enriching visual knowledge bases via object discovery and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2027–2034.

[13] S. Vicente, J. Carreira, L. Agapito, and J. Batista, "Reconstructing PASCAL VOC," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 41–48.

[14] D. Zhang, J. Han, Y. Yang, and D. Huang, "Learning category-specific 3D shape models from weakly labeled 2D images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3587–3595.

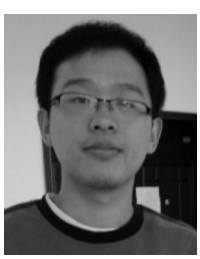
[15] L. Torresani, A. Hertzmann, and C. Bregler, "Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 5, pp. 878–892, May 2008.

[16] S. Satkin, M. Rashid, J. Lin, and M. Hebert, "3DNN: 3D nearest neighbor," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 69–97, 2015.

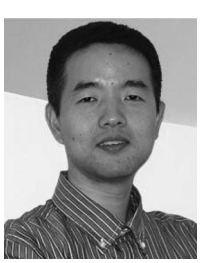
[17] J. J. Lim, H. Pirsiavash, and A. Torralba, "Parsing IKEA objects: Fine pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2992–2999.
[18] Z. Wu et al., "3D shapenets: A deep representation for volumetric shapes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1912–1920.
[19] 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 *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 628–644.
[20] D. J. Rezende, S. A. Eslami, S. Mohamed, P. Battaglia, M. Jaderberg, and N. Heess, "Unsupervised learning of 3D structure from images," in *Proc. Neural Inf. Process. Syst.*, 2016, pp. 4996–5004.
[21] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proc. 26th Annu. Conf. Comput. Graphics Interactive Tech.*, 1999, pp. 187–194.
[22] M. Z. Zia, M. Stark, B. Schiele, and K. Schindler, "Detailed 3D representations for object recognition and modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2608–2623, Nov. 2013.
[23] S. Yingze Bao, M. Chandraker, Y. Lin, and S. Savarese, "Dense object reconstruction with semantic priors," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1264–1271.
[24] A. Dame, V. A. Prisacariu, C. Y. Ren, and I. Reid, "Dense reconstruction using 3D object shape priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1288–1295.
[25] S. Zhu, L. Zhang, and B. M. Smith, "Model evolution: An incremental approach to non-rigid structure from motion," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1165–1172.
[26] M. Prasad, A. Fitzgibbon, A. Zisserman, and L. Van Gool, "Finding nemo: Deformable object class modelling using curve matching," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1720–1727.
[27] C. Bregler, A. Hertzmann, and H. Biermann, "Recovering non-rigid 3D shape from image streams," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2000, pp. 690–696.
[28] M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *Proc. 23rd Int. Conf. Neural Inf. Process. Syst.*, vol. 1, 2010, pp. 1189–1197.
[29] T. Schuster, L. Wolf, and D. Gadot, "Optical flow requires multiple strategies (but only one network)," in *Proc. Comput. Vis. Pattern Recognit.*, 2017, pp. 4950–4959.
[30] B. Hariharan, J. Malik, and D. Ramanan, "Discriminative decorrelation for clustering and classification," *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 459–472.
[31] S. Vicente, V. Kolmogorov, and C. Rother, "Graph cut based image segmentation with connectivity priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.
[32] D. Zhang, J. Han, L. Zhao, and D. Meng, "Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework," *Int. J. Comput. Vis.*, vol. 127, no. 4, pp. 363–380, 2019.
[33] D. Zhang, J. Han, L. Yang, and D. Xu, "SPFTN: A joint learning framework for localizing and segmenting objects in weakly labeled videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Nov. 13, 2018, doi: 10.1109/TPAMI.2018.2881114.
[34] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
[35] N. Aspert, D. Santa-Cruz, and T. Ebrahimi, "Mesh: Measuring errors between surfaces using the hausdorff distance," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2002, pp. 705–708.
[36] Y. Xiang, R. Mottaghi, and S. Savarese, "Beyond PASCAL: A benchmark for 3D object detection in the wild," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2014, pp. 75–82.
[37] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.
[38] M. Paladini, A. Del Bue, J. Xavier, L. Agapito, M. Stošić, and M. Dodig, "Optimal metric projections for deformable and articulated structure-from-motion," *Int. J. Comput. Vis.*, vol. 96, no. 2, pp. 252–276, 2012.
[39] Y. Dai, H. Li, and M. He, "A simple prior-free method for non-rigid structure-from-motion factorization," *Int. J. Comput. Vis.*, vol. 107, no. 2, pp. 101–122, 2014.
[40] A. Eriksson and A. Van Den Hengel, "Efficient computation of robust low-rank matrix approximations in the presence of missing data using the L1 norm," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 771–778.
[41] M. Marques and J. Costeira, "Estimating 3D shape from degenerate sequences with missing data," *Comput. Vis. Image Understanding*, vol. 113, no. 2, pp. 261–272, 2009.
[42] N. R. Twarog, M. F. Tappen, and E. H. Adelson, "Playing with puffball: Simple scale-invariant inflation for use in vision and graphics," in *Proc. ACM Symp. Appl. Perception*, 2012, pp. 47–54.
[43] J. T. Barron and J. Malik, "Color constancy, intrinsic images, and shape estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 57–70.
[44] A. Kanazawa, D. W. Jacobs, and M. Chandraker, "WarpNet: Weakly supervised matching for single-view reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3253–3261.
[45] A. Joulin, F. Bach, and J. Ponce, "Discriminative clustering for image co-segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1943–1950.
[46] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1377–1385.

Junwei Han is a currently a full professor with Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, multimedia processing, and brain imaging analysis. He is an associate editor of the *IEEE Transactions on Human-Machine Systems*, *Neurocomputing*, and *Multidimensional Systems and Signal Processing*. He is a senior member of the IEEE.

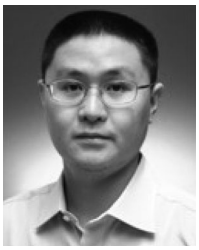
Yang Yang received the BE degree from the Shaanxi University of Science and Technology, Xi'an, China, in 2010, and the MS degree from Lanzhou University of Technology, Lanzhou, China. He is currently working toward the PhD degree at Northwestern Polytechnical University, Xi'an. His research interests include 3D vision, and weakly supervised learning.



Dingwen Zhang received the BE and PhD degrees from the Northwestern Polytechnical University, Xi'an, China, in 2012 and 2018, respectively. From 2015 to 2017, he was a visiting scholar with Carnegie Mellon University, Pittsburgh, PA. He is currently an associate professor with Xidian University, Xi'an. His research interests include computer vision and multimedia processing, especially on saliency detection, co-saliency detection, and weakly supervised learning.



Dong Huang received the MSc degree in automation and the PhD degree in computer science from the University of Electronic Science and Technology of China, Chengdu, China in 2005 and 2009, respectively. He is current a project scientist with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA. His research focuses on computer vision and machine learning.



Dong Xu (M'07-SM'13) is currently a professor (chair in computer engineering) with the School of Electrical and Information Engineering, The University of Sydney, Sydney, NSW, Australia. He has published more than 100 papers in the IEEE TRANSACTIONS and top tier conferences. He co-authored work on transfer learning for video event recognition received the Best Student Paper Award in the IEEE International Conference on Computer Vision and Pattern Recognition in 2010. He is a fellow of the IEEE and IAPR.

Fernando de la Torre received the BSc degree in telecommunications, the MSc and PhD degrees in electronic engineering from the La Salle School of Engineering, Ramon Llull University, Barcelona, Spain, in 1994, 1996, and 2002, respectively. In 2003, he joined the Robotics Institute, Carnegie Mellon University and he is currently an associate research professor. His research interests include the fields of computer vision and machine learning.