Weakly-supervised Reconstruction of 3D Objects with Large Shape Variation from Single In-the-Wild Images

Shichen Sun[1], Zhengbang Zhu[1], Xiaowei Dai[1], Qijun Zhao[1,2,*], and Jing Li[1]

[1] College of Computer Science, Sichuan University, China
[2] School of Information Science and Technology, Tibet University, China

Abstract. Existing unsupervised 3D object reconstruction methods can not work well if the shape of objects varies substantially across images or if the images have distracting background. This paper proposes a novel learning framework for reconstructing 3D objects with large shape variation from single in-the-wild images. Considering that shape variation leads to appearance change of objects at various scales, we propose a fusion module to form combined multi-scale image features for 3D reconstruction. To deal with the ambiguity caused by shape variation, we propose side-output mask constraint to supervise the feature extraction, and adaptive edge constraint and initial shape constraint to supervise the shape reconstruction. Moreover, we propose background manipulation to augment the training images such that the obtained model is robust to background distraction. Extensive experiments have been done for both non-rigid objects (birds) and rigid objects (planes and vehicles), and the results prove the superiority of the proposed method.

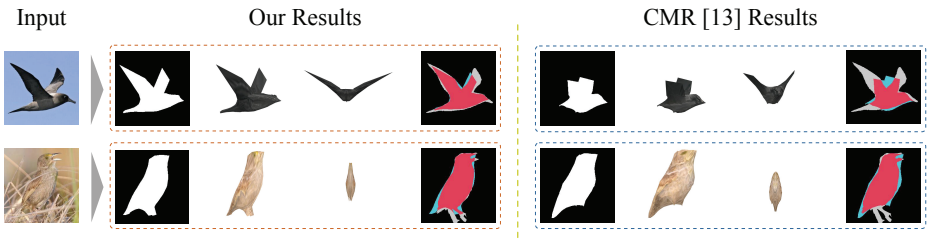

Fig. 1: Our proposed 3D object reconstruction method can infer more accurate 3D shapes especially for objects with large shape variation and images with cluttered background. For easy comparison, we overlay ground truth masks (in grey color) with the reconstructed ones (in cyan color), and highlight their common regions in red color.

1 Introduction

Knowing 3D shapes of objects is of significant importance in many tasks, e.g., scene understanding [1] and surgical navigation. While inferring the 3D shape of an object from a single view of the object seems effortless for the human vision system, it is still quite difficult for computer vision systems. Some researchers implement 3D object reconstruction by using multi-view geometry based approaches [2][3][4], which estimate 3D object shapes by exploring the motion or appearance clues among the multiple views of objects. These methods are limited by the availability of multi-view images of objects, and are consequently not always applicable in different scenarios. In this paper, we focus on 3D object reconstruction from single images.

Learning-based approaches are nowadays dominant in 3D object reconstruction. These methods use 3D volumes [5], point clouds [6], or meshes [7] to represent 3D objects, among which meshes can provide finer shape details and effectively support various shape editing and are thus preferred in many applications. Most of these methods [8][9][10] require ground truth 3D shapes as supervision during learning. Some of them [11] first learn morphable models based on the ground truth 3D shapes and then estimate 3D shapes for input images by fitting the morphable models. Others [12] learn mappings from 2D images to 3D shape deformations that are needed to transform the initial shape estimate towards its true value. Despite the impressive results obtained by these methods, collecting ground truth 3D shapes of objects is not always affordable or feasible. As a result, learning for 3D object reconstruction without supervision of ground truth 3D shapes is attracting increasing attention.

When ground truth 3D shapes are not available for the training 2D images, 3D object reconstruction learning can be supervised by various prior-based constraints or by features on 2D images. While priors such as smoothness [13] and low-rank [14] have been proven effective for unsupervised 3D object reconstruction, features like keypoints, silhouettes, foreground masks, texture values, and perceptual features [15] are widely used to enforce the coincidence between the input image and the image rendered from the estimated 3D object. To apply such supervision, existing methods [16] simply extract from the input image a latent feature representation, which is assumed to encode the shape, texture and camera parameters associated with the input image, and based on which the 3D object in the image is regressed. Despite the impressive results obtained by them, they work poorly when the objects have large shape variation or when the images have distracting background, as shown in Fig. **??**. We believe that this is due to their inefficient utilization of data (or features) and constraints. For instance, when enforcing the coincidence, existing methods consider only locations of keypoints and apply geometric constraints like foreground mask constraint at the output layers only. As a consequence, they are poor both at dealing with invisible keypoints that are caused either by occlusion or by large pose variations of objects and at capturing rich shape deformations of objects.

This paper aims to boost the 3D reconstruction performance for objects with large shape variation and for images with cluttered background. To this

end, we propose a novel learning framework by exploiting the training data and the geometric constraints in more efficient manners. Specifically, considering that shape variation leads to appearance change of objects at various scales, we propose a fusion module to combine multi-scale features extracted from the input image, based on which the 3D object is estimated. Considering the ill-posed nature of reconstructing 3D objects from single 2D images and the ambiguity caused by shape variation, we propose side-output mask constraint to supervise the feature extraction, and adaptive edge constraint and initial shape constraint to supervise the shape reconstruction. Moreover, we augment the training images via manipulating their background to improve the robustness of the obtained model to background distraction. We validate the superiority of our proposed method by extensive experiments for reconstructing both non-rigid objects, i.e., birds, and rigid objects, i.e., vehicles and planes.

2 Related Work

In this section, we first discuss the shape representations used in 3D object reconstruction, and then review the 3D reconstruction methods for animals, typical categories of objects with large shape variation.

Shape Representations: Early deep-learning-based methods [17][18] directly predict the final output shapes using voxel in low-resolution due to the high computation cost of 3D convolution operators. Based on the fact that the core difference between high-resolution and low-resolution shapes lies in the boundary surface details and detailed shape information, methods in [10][19] take octree, a sub-category of volume, as representation to implement high-resolution reconstruction in a computationally-efficient manner. Compared to volume, point cloud represents 3D shapes in a more flexible and expressive way. Fan *et al.* [20] firstly propose a framework to generate 3D shapes based on point cloud. Many other methods [21][22][23][24] then choose point cloud as representation and focus on how to alleviate the shape ambiguity and improve accuracy. Nevertheless, the predicted point clouds are still of low accuracy. Consequently, more and more methods [25][3][16][26][7][9][12][27] have replaced point cloud with mesh as the representation of 3D shapes. Mesh, particularly triangular mesh composed of node and triangular face, describes shapes in a more comprehensive way, enabling not only topology constraints but also alignment between shapes. Therefore, we will use mesh to represent 3D objects in this paper.

3D Animal Reconstruction: Little work has been done on 3D animal reconstruction. The seminal work in [28] learns a deformable model of animals from several images. The method is however restricted by precise manual annotations and not ready for strongly articulated objects. Method in [29] takes a set of segmented images as input, and adopts a patch-based approach to implement reconstruction. This way, articulated and relatively accurate 3D animal shapes can be reconstructed. In order to model 3D animal shape as a whole, method in [30] captures shape deformation by defining the stiffness value of local regions. To ease the lack of 3D scans, method in [31] instead uses 3D scans of toy an-

imals to learn a parametric model called SMAL. Later, method in [32] makes the initial parametric model fit the characteristic of the individual shape of the given animal before optimization such that some species unseen in training set could have a better reconstruction. Most recently, a learning-based approach called SMALST [33] integrates SMAL model into a regression network. The method uses the existing SMAL model to spawn training data in various poses, shapes, camera parameters and backgrounds, which are naturally equipped with ground truth 2D annotations. Without relying on parametric model, Kanazawa *et al.* [13] learn from a collection of images of a specific category of objects (e.g., birds) a regression network that can deform an initial shape toward the true shape of the object instance in the input image. Common drawbacks of these methods include (i) exploiting features in a coarse manner without considering the impact of shape variation on object appearance, (ii) inefficient utilization of geometric constraints resulting in ambiguity in the reconstructed shape and low coincidence between the obtained 3D object and the input image, and (iii) poor generalization to in-the-wild images with messy background. As we will show in this paper, our proposed method can effectively get rid of these drawbacks and substantially improve the 3D object reconstruction accuracy.

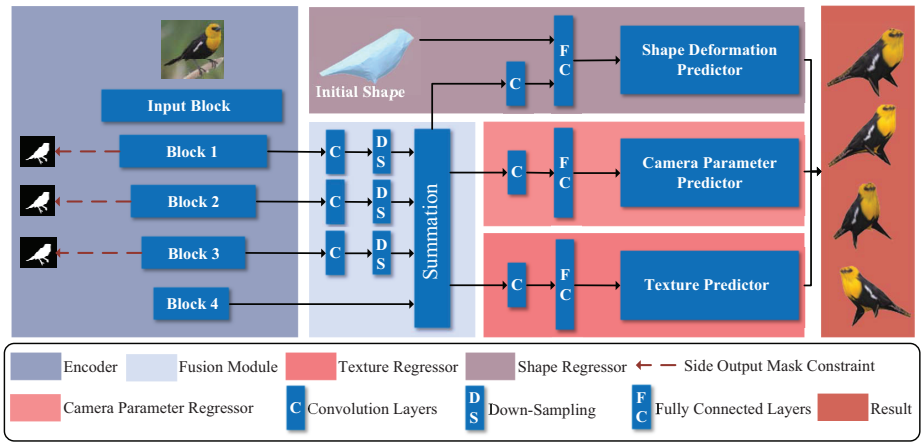

Fig. 2: Schematic diagram of the proposed method of unsupervised 3D object reconstruction from single images.

3 Method

3.1 Overview

Fig. 2 shows the overall framework of our proposed single-image-based unsupervised 3D object reconstruction method. The input is a single RGB image

of an instance of the target object category (e.g., birds). Note that during test the method does not require any annotation. To reconstruct the 3D model of the object instance, shape deformation, in addition to the UV-flow (i.e., texture feature; see ref. [13] for detail) and camera parameters, is estimated with respect to an initial shape. First, latent representations at multiple scales are extracted by using a backbone deep encoder network, and combined by using a fusion module. Shape deformation, UV-flow and camera parameters are then predicted all by inference from the fused features but with respective regressors. Multi-scale features are used such that richer appearance change induced by shape variation can be captured in 3D reconstruction. It is worth highlighting that the initial shape is also taken as input by the shape deformation regressor, which serves as additional constraint on the search of correct shape deformation. The 3D shape of the object in the input image can be finally obtained by applying the estimated shape deformation to the initial shape, while the texture value of each vertex in the 3D object shape can be obtained according to the estimated UV-flow and the input RGB image.

To determine the parameters involved in the encoder and regressor networks, a set of images of the target category of objects is required as training data. These training images are annotated with keypoints and foreground masks, but do not have ground truth 3D shapes. The training is supervised based on the re-projection of the reconstructed 3D object onto 2D image plane and by coincidence constraints and various prior constraints. In the rest of this section, we introduce in detail the key components of our proposed method for coping with shape variation and cluttered image background, including inference with multi-scale features, shape-sensitive geometric constraints, and training data augmentation via background manipulation.

3.2 Inference with Multi-Scale Features

Existing 3D object reconstruction methods mostly extract latent representation at a single scale, i.e., the feature of the deepest layer of the encoder, based on which shape deformation, texture, and camera parameters all are predicted. In contrast, many other tasks have proven the necessity of using features at multiple scales. Specific to 3D object reconstruction, especially for objects with large shape variation, we argue that it is important to fuse multi-scale features for 3D object inference because shape variation could lead to object appearance change at a variety of scales. We thus propose to fuse the multi-scale features extracted from different layers of the encoder network for predicting 3D objects.

Unlike the method in [34] that concatenates multiple features to estimate 3D objects, we propose a convolution-based fusion module to combine multi-scale features. This is because the concatenation method dramatically increases the number of parameters which makes the network difficult to converge, especially without 3D supervision in our case. To solve this problem, in our proposed fusion module, the feature maps of lower layers are first convoluted and down-sampled to the same size as the feature map of higher layer, and afterwards the feature maps of different layers are added up via element-wise summation to produce

the fused feature. Obviously, our proposed fusion module, while itself has very few parameters, does not affect the complexity of the regression network.

3.3 Shape-Sensitive Geometric Constraints

Keypoints, silhouette, and foreground mask are routinely used by existing unsupervised 3D object reconstruction methods as geometric constraints by enforcing the coincidence of these geometric features between the input image and the image generated from the reconstructed 3D object. Yet, as we will show below, the way of existing methods to apply geometric constraints can not effectively supervise the extraction of shape-sensitive features or avoid shape ambiguity. Therefore, we propose the following three shape-sensitive geometric constraints to enhance the ability of learned 3D object reconstruction model to handle objects with large shape variation.

Side output mask constraint. While existing methods train the 3D object reconstruction network in an end-to-end manner, the geometric constraints applied on the final output might be of low efficiency in supervising the training of the encoder that is located at the frontal end of the entire network. Moreover, as being discussed above, multi-scale features extracted by different layers of the encoder are employed to regress the 3D object. Therefore, it is demanded to make the features capture as much shape information as possible. To this end, we propose to directly predict foreground mask from the feature extracted by the intermediate layer of the encoder. We use these side output foreground masks to evaluate the mask coincidence as additional supervision for training the encoder, which is defined as follows.

$$\begin{aligned} L_{sidemask} = & \sum_{i=1}^{N} \sum_{k=1}^{N_b} \sum_{(x,y)} [-M_{gt,i}(x,y) \log(M_{pred,i}^{k}(x,y)) \\ & - (1 - M_{gt,i}(x,y)) \log(1 - M_{pred,i}^{k}(x,y))], \end{aligned} \tag{1}$$

where $M_{gt,i}$ and $M_{pred,i}$ denote the ground truth mask and predicted mask, respectively, N is the total number of training images, N_b is the number of intermediate blocks that are considered in side output mask constraint, and (x, y) denotes the pixel on the mask.

Edge constraint. Using mean squared error (MSE) to measure the coincidence of keypoints, existing methods face the difficulty in coping with ambiguous shapes. As shown in Fig. 3, two predictions of the four keypoints share the same MSE, but they are obviously not of the same optimality with respect to the ground truth. This is partially due to the missing edge or inter-keypoint constraint that can capture the shape topology. Motivated by the recent progress in human body pose estimation [35], we define the following edge loss to enforce the topology coincidence of the keypoints.

$$L_{edge} = \sum_{i=1}^{N} \sum_{j=1}^{N_E^i} \|E_{i,j} - \hat{E}_{i,j}\|_2^2, \quad E_{i,j} \in \mathcal{DT}(X_i), \tag{2}$$

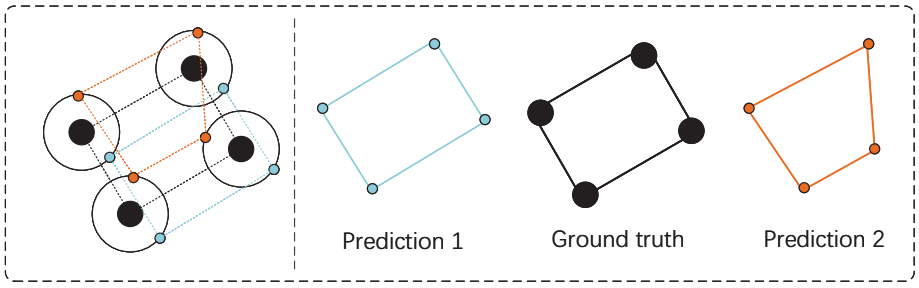

Fig. 3: Existing unsupervised 3D object reconstruction methods use mean squared error (MSE) to measure the coincidence of keypoints. MSE can not deal with ambiguous shapes: The two predictions share the same MSE; however, Prediction 1 is obviously better than Prediction 2. Introducing edge constraint can alleviate this problem.

where E and $\hat{E}$ are, respectively, the ground truth and estimated edges connecting two keypoints, $\mathcal{DT}(*)$ denotes the operation creating the edge set for a set of visible keypoints via Delaunay Triangulation, i refers to the $i^{\texttt{th}}$ of the N training images, j refers to the $j^{\texttt{th}}$ of the N_E^i edges on the $i^{\texttt{th}}$ image, and X_i is the set of visible 2D keypoints on the $i^{\texttt{th}}$ image. Different from the edge loss defined in human body pose estimation which uses pre-specified keypoint connectivity, we adaptively generate edges for each image based on the visible keypoints on it. This way, our method can effectively deal with objects with large pose variation.

Initial shape constraint. Although existing unsupervised 3D object reconstruction methods mostly infer shape deformation (with respect to an initial shape estimate) rather than the shape itself, they calculate the shape deformation based purely on the extracted feature of the input image without considering the initial shape at all. We believe that explicitly utilizing the initial shape when predicting the shape deformation can serve as another constraint and thus help to improve the prediction accuracy. For this sake, we propose to concatenate the vertex coordinates of the initial shape with the extracted feature and feed the expanded feature vector into the shape deformation regressor. This is effective especially for largely deformed objects, e.g., birds with open wings.

3.4 Training Data Augmentation via Background Manipulation

In this paper, we employ a pre-trained ResNet-18 [36] as the backbone encoder network, and train the entire 3D object reconstruction network in an end-to-end manner with the following overall loss function,

$$\begin{aligned} Loss = &\lambda_1 L_{edge} + \lambda_2 L_{sidemask} + \lambda_3 L_{kp} + \lambda_4 L_{sil} \\ &+ \lambda_5 L_{cam} + \lambda_6 L_{smooth} + \lambda_7 L_{def} + \lambda_8 L_{texture}. \end{aligned} \tag{3}$$

Here, L_{kp}, L_{sil}, and L_{cam} are defined as the MSE loss between ground truth and predicted values of keypoints, silhouettes, and camera parameters, respectively.

The predicted 2D key-points are computed by reprojecting 3D key-points back to the image plane. We use NMR [37] to get the predicted silhouettes and texture under the observed view. L_{smooth} and L_{def} as defined in [13] are used to constrain the inferred shape to be natural; specifically, L_{smooth} is the discrete Laplace-Beltrami operator and L_{def} is to penalize large deformation. $L_{texture}$, a perceptual loss [38], is used to constrain the prediction of texture.

As we attempt to apply our proposed method to reconstructing 3D objects from in-the-wild images, we have to consider the impact of clutter background. Taking bird images as example, we observe that foreground birds could appear quite similar to the background in real-world images because of the natural camouflage of birds. Such camouflage would distract the trained reconstructor during testing as the foreground is not annotated on the test image. To solve this problem, we propose to augment the training data with images that are generated from the original training images by erasing or substituting the background (note that foreground has been annotated on the training images). We will experimentally show the effectiveness of such augmentation though it is very simple to apply.

4 Experiments

4.1 Implementation Details

The network is implemented in Pytorch and optimized using Adam with batch size as 16 and learning rate as 1e-5. The values of hyperparameters in Equation (3) are set as $\lambda_1 = 5.0$, $\lambda_2 = 5.0$, $\lambda_3 = 60.0$, $\lambda_4 = 5.0$, $\lambda_5 = 5.0$, $\lambda_6 = 50.0$, $\lambda_7 = 10.0$, $\lambda_8 = 0.5$. We assume that objects are at the center of images. In our experiments, we crop the images according to the bounding boxes of objects such that the objects locate at image center.

4.2 Datasets and Protocols

We evaluate our proposed method with comparison to state-of-the-art (SOTA) methods for reconstructing both non-rigid objects and rigid objects. For non-rigid objects, we take birds as the target 3D objects, and use the CUB-200-2011 dataset. For the sake of fair comparison with the SOTA bird reconstruction method, namely CMR, in [13], we follow the same setup of data division into training, validation, and test sets. Each bird image is annotated with $9 \sim 15$ keypoints, foreground mask as well as camera parameters. As for rigid objects, we consider vehicles and planes, and use the PASCAL 3D+ dataset [39] with the same data division as for the counterpart methods.

Due to the lack of ground truth 3D shapes, as being normally done in the literature [13][15], we also use the two metrics of Intersection over Union (IoU) and Percentage of Correct Keypoints (PCK) to assess the reconstruction accuracy. However, IoU puts more weight on the interior of reconstructed objects, while neglecting to some extent the discrepancy of boundary. For a more comprehensive evaluation of the reconstruction performance, therefore, we propose to

use the structural similarity (SSIM) as another metric to measure the similarity between the input image and the rendered image. SSIM as a perceptual metric can effectively measure the structural difference, which is essential in evaluating the reconstructed 3D objects.

In the following experiments, we conduct model analysis and ablation study on the CUB-200-2011 dataset, and compare the proposed method with state-of-the-art methods on both CUB-200-2011 dataset and PASCAL 3D+ dataset.

4.3 Model Analysis

We compare different implementations of the proposed method to evaluate the impact of (i) number of side output mask constraints, (ii) definition of edges in edge constraint, (iii) background manipulation methods in data augmentation, and (iv) down-sampling methods in feature fusion.

Number of side output mask constraints. We consider four cases $\{S_i|i = 1, 2, 3, 4\}$ where S_i denotes applying side output mask constraint for the first i blocks following the input block (ordered from shallow to deep) in the encoder network. Fig. 4 presents the mean IoU with regard to the number of side output mask constraints. The results demonstrate that more side output mask constraints can generally improve the reconstruction accuracy, but the performance gain becomes saturated as more deep blocks are included.

Definition of edges. In this experiment, besides our proposed edge definition (DT), we consider three other definitions of edges, i.e., two prior-knowledge-based manual definitions (M1 and M2) and the full set of edges between visible keypoints (FC). The results are shown in Table 1. As can be seen, edge loss can effectively improve the reconstruction accuracy, and defining edges adaptively according to visible keypoints is better than using fixed edge definitions. Moreover, considering the computational cost, the proposed edge definition is more preferred than using the full set of edges.

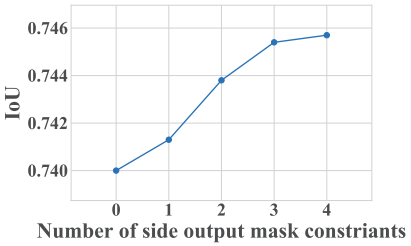

Fig. 4: Impact of number of side output mask constraints.

Background manipulation methods. Table 2 shows the results when different background manipulation methods are applied for data augmentation. We can see that while manipulating background of training images is effective in improving the reconstruction accuracy, the best way is to substituting the background pixel values with the average values of the background pixels across all the training images.

Down-sampling methods. In this experiment, we implement and compare three down-sampling methods for feature fusion: sampling (choose the value of center pixel in the sampling grid), average pooling, and max pooling. Table 3 gives the results. Note that according to the above evaluation results, we apply

Table 1: Impact of edge definitions. DT is the proposed one. M1 and M2 are two manual definitions (see supplementary material for detail) that are fixed for all images no matter which keypoints are visible. FC denotes the full set of edges between visible keypoints. Time shows the additional time required for evaluating the edge loss.

Method	Time	IoU	PCK.1	PCK.15	SSIM
Baseline	-	0.740	0.783	0.916	0.8568
Baseline + edge (DT)	**0.01062**	0.747	0.848	0.943	**0.8587**
Baseline + edge (M1)	0.02320	0.742	0.813	0.922	0.8574
Baseline + edge (M2)	0.05936	0.745	0.828	0.938	0.8580
Baseline + edge (FC)	0.12469	**0.748**	**0.855**	**0.952**	0.8581

Table 2: Impact of background manipulation methods in data augmentation. Black, White, and Mean denote replacing the background pixel values with black $(0, 0, 0)$, white $(255, 255, 255)$, and the average values of the background pixels across all the training images, respectively.

Method	IoU	PCK.1	PCK.15	SSIM
Baseline	0.740	0.783	0.916	0.8568
Baseline + Black	0.743	0.854	0.951	0.8573
Baseline + White	0.750	0.852	0.949	0.8601
Baseline + Mean	**0.753**	**0.855**	**0.952**	**0.8605**

the edge constraint ('edge'), side output mask constraint ('so') as well as initial shape constraint ('isc') to the baseline in this experiment. As can be seen, the reconstruction accuracy is further improved after feature fusion. Moreover, among the three down-sampling methods, max pooling achieves the best results. We argue that max pooling has the ability to preserve the prominent feature benefiting the reconstruction, whereas average pooling would distract the regressor from the prominent feature, leading to poor results.

4.4 Ablation Study

In this experiment, we evaluate the contribution of different components to the performance gain of the proposed method. For this purpose, we gradually integrate the following components: initial shape constraint ('isc'), edge constraint ('edge'), side output mask constraint ('so') and multi-scale feature fusion ('msf'). Note that in this experiment the best implementation for 'edge', 'so' and 'msf' is employed according to the model analysis results. Table 4 summarizes the ablation study results, which clearly demonstrate that all the proposed components effectively improve the reconstruction accuracy.

4.5 Comparison to State-of-the-arts

Table 3: Impact of down-sampling methods in feature fusion. Three down-sampling stratagies, Sampling at center pixels (Sampling), Average pooling (AvgPool) and Max pooling (MaxPool), are implemented.

Method	IoU	PCK.1	PCK.15	SSIM
Baseline	0.740	0.783	0.916	0.8568
Baseline + isc	0.742	0.798	0.933	0.8572
Baseline + isc + edge + so	0.749	0.851	0.953	0.8597
Baseline + isc + edge + so + Sampling	0.756	0.862	0.954	0.8620
Baseline + isc + edge + so + Avgpool	0.754	0.858	0.952	0.8607
Baseline + isc + edge + so + Maxpool	**0.757**	**0.866**	**0.957**	**0.8631**

Table 4: Ablation study results. 'isc', 'edge', 'so' and 'msf' denote, respectively, initial shape constraint, edge constraint, side output mask constraint and multi-scale feature fusion.

Model	IoU	PCK.1	PCK.15	SSIM
Baseline	0.740	0.783	0.916	0.8568
Baseline + isc	0.742	0.798	0.933	0.8572
Baseline + isc + edge + so	0.749	0.851	0.953	0.8597
Baseline + isc + edge + so + msf	**0.757**	**0.866**	**0.957**	**0.8631**

We lastly compare our method with the state-of-the-art (SOTA) method CMR [13] on CUB-200-2011 dataset. The overall results in terms of PCK and IoU are presented in Fig. 5. It can be observed that the proposed method consistently outperforms CMR. The average results in terms of different metrics are shown in Table 5. Some examples of reconstructed 3D birds are shown in Fig. 6. As can be seen, our proposed method can generate visually more pleasant shapes, particularly for the torso and wings of birds.

We also evaluate our method for the reconstruction of vehicles and planes. We use the images in PASCAL VOC [39] and ImageNet [13] for training. An off-the-shelf segmentation framework [40] is used to obtain the silhouettes (and thus foreground masks) for the images. We report the IoU re-

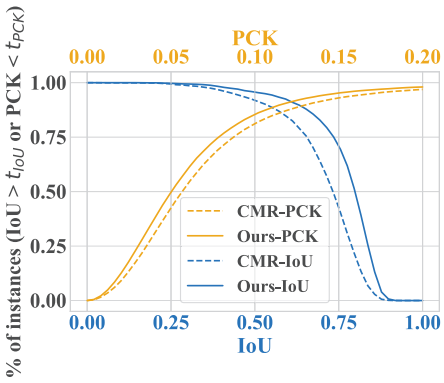

Fig. 5: Comparison with the SOTA method CMR [13] on CUB-200-2011. X-axis represents threshold (t_{PCK} or t_{IoU}) and Y-axis is the proportion of test instances whose PCK/IoU is lower/higher than the threshold.

sults of our method and the counterpart methods on the test PASCAL 3D+ dataset in Table 6. Our method improves the reconstruction accuracy for both categories of objects with a large margin. Some reconstructed planes and vehicles are visualized in Fig. 7.

Table 5: Comparison with the state-of-the-art method CMR [13] for 3D bird reconstruction on CUB-200-2011.

Method	IoU	PCK.1	PCK.15	SSIM
CMR	0.703	0.812	0.93	0.8439
Ours	**0.766**	**0.854**	**0.953**	**0.8657**

Table 6: Comparison of our method with state-of-the-art methods for reconstructing vehicles and planes on the PASCAL 3D+ dataset [39] in terms of IoU. '+pose' indicates that the method requires ground truth camera parameters as input during test.

Method	Planes	Vehicles
CSDM [41]	0.398	0.600
DRC + pose [42]	0.415	0.666
CMR [13]	0.460	0.640
VPL + pose [15]	0.475	0.679
Ours	**0.584**	**0.853**

5 Conclusion

In this paper, we have made an attempt to boost the accuracy of reconstructing 3D objects with large shape variation from single in-the-wild images when 3D supervision is not available during training. Specifically, it provides an efficient and effective fusion module for aggregating multi-scale features for 3D reconstruction, and trains the entire reconstruction network with shape-sensitive geometric constraints including edge constraint, side output mask constraint, and initial shape constraint. Moreover, by augmenting the training data via manipulating the background in training images, our method can better deal with real-world images with distracting background. The effectiveness of our method has been proven on images of birds, vehicles and planes.

6 Acknowledgments

This work is supported by the National Natural Science Foundation of China (61773270, 61971005).

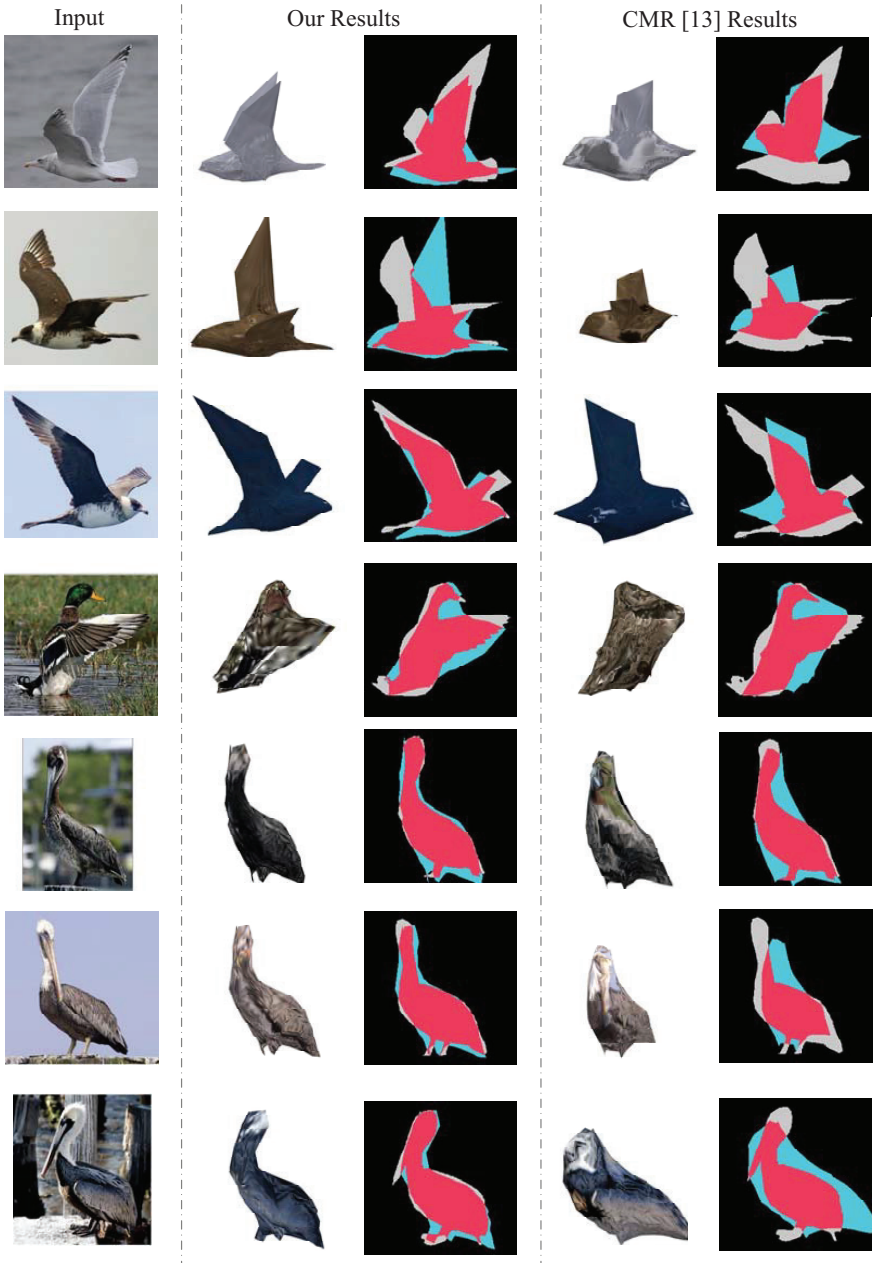

Fig. 6: Reconstruction results of our method and the CMR method [13] on CUB-200-2011. For easy comparison, we overlay ground truth masks (in grey color) with the reconstructed ones (in cyan color), and highlight their common regions in red color.

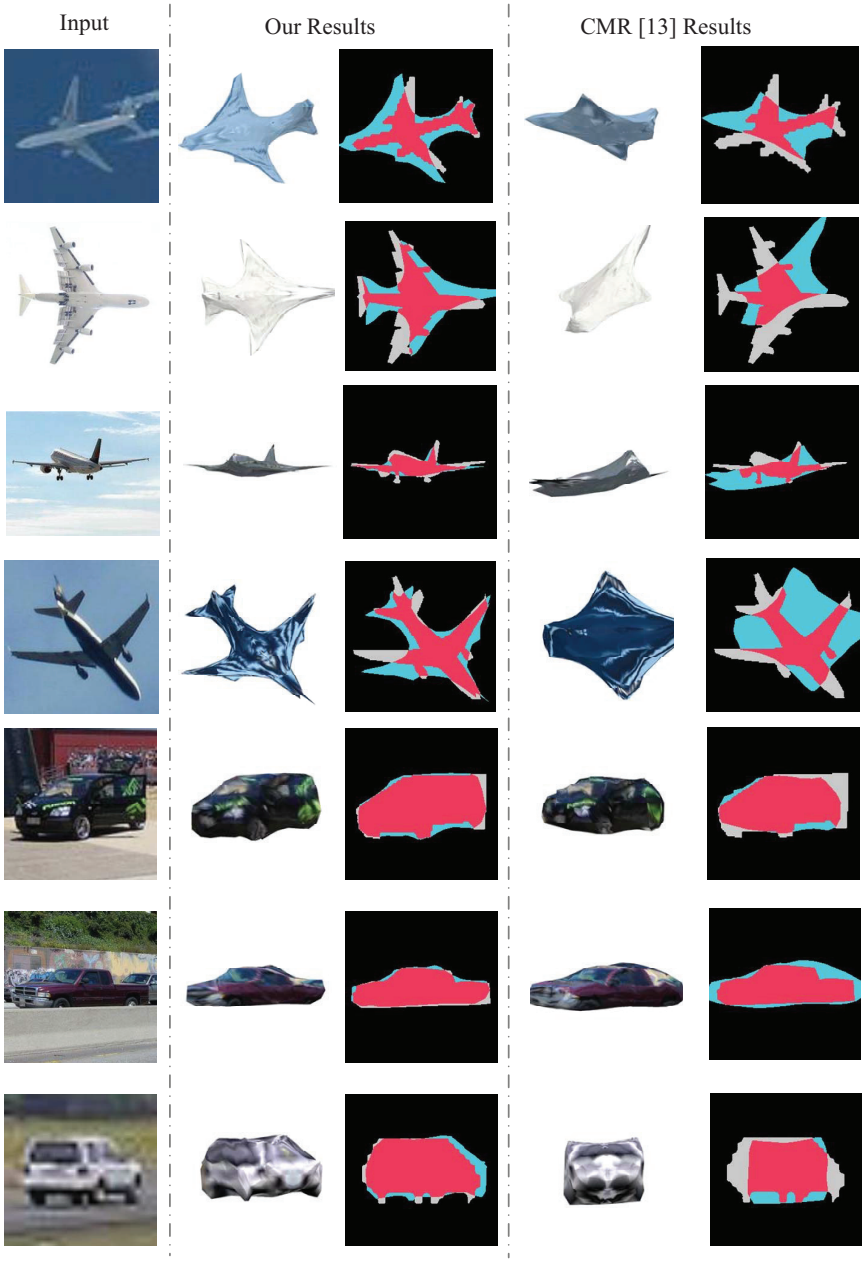

Fig. 7: Reconstruction results of our method and the CMR method [13] on PASCAL 3D+. For easy comparison, we overlay ground truth masks (in grey color) with the reconstructed ones (in cyan color), and highlight their common regions in red color.

References

1. Zhang, P., Liu, W., Lei, Y., Lu, H., Yang, X.: Cascaded context pyramid for full-resolution 3d semantic scene completion. In: IEEE International Conference on Computer Vision (ICCV). (2019) 7801–7810
2. Kar, A., Häne, C., Malik, J.: Learning a multi-view stereo machine. In Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., eds.: Advances in Neural Information Processing Systems (NeurIPS). (2017) 365–376
3. Lin, C.H., Wang, O., Russell, B.C., Shechtman, E., Kim, V.G., Fisher, M., Lucey, S.: Photometric mesh optimization for video-aligned 3d object reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019) 969–978
4. Sridhar, S., Rempe, D., Valentin, J., Bouaziz, S., Guibas, L.J.: Multiview aggregation for learning category-specific shape reconstruction. In: Advances in Neural Information Processing Systems (NeurIPS). (2019) 2348–2359
5. Shen, W., Jia, Y., Wu, Y.: 3d shape reconstruction from images in the frequency domain. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019) 4471–4479
6. Yang, G., Huang, X., Hao, Z., Liu, M.Y., Belongie, S., Hariharan, B.: Pointflow: 3d point cloud generation with continuous normalizing flows. In: IEEE International Conference on Computer Vision (ICCV). (2019) 4541–4550
7. Pan, J., Han, X., Chen, W., Tang, J., Jia, K.: Deep mesh reconstruction from single rgb images via topology modification networks. In: IEEE International Conference on Computer Vision (ICCV). (2019) 9964–9973
8. Richter, S.R., Roth, S.: Matryoshka networks: Predicting 3d geometry via nested shape layers. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018) 1936–1944
9. Smith, E., Fujimoto, S., Romero, A., Meger, D.: Geometrics: Exploiting geometric structure for graph-encoded objects. In Chaudhuri, K., Salakhutdinov, R., eds.: International Conference on Machine Learning (ICML). (2019) 5866–5876
10. Tatarchenko, M., Dosovitskiy, A., Brox, T.: Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In: IEEE International Conference on Computer Vision (ICCV). (2017) 2088–2096
11. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH). (1999) 187–194
12. Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., Jiang, Y.G.: Pixel2mesh: Generating 3d mesh models from single rgb images. In: European Conference on Computer Vision (ECCV). (2018) 52–67
13. Kanazawa, A., Tulsiani, S., Efros, A.A., Malik, J.: Learning category-specific mesh reconstruction from image collections. In: European Conference on Computer Vision (ECCV). (2018) 371–386
14. Cha, G., Lee, M., Oh, S.: Unsupervised 3d reconstruction networks. In: The IEEE International Conference on Computer Vision (ICCV). (2019) 3849–3858
15. Kato, H., Harada, T.: Learning view priors for single-view 3d reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019) 9778–9787
16. Liu, S., Saito, S., Chen, W., Li, H.: Learning to infer implicit surfaces without 3d supervision. In: Advances in Neural Information Processing Systems (NeurIPS). (2019) 8293–8304

17. Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In: European Conference on Computer Vision (ECCV). (2016) 628–644
18. Girdhar, R., Fouhey, D., Rodriguez, M., Gupta, A.: Learning a predictable and generative vector representation for objects. In: European Conference on Computer Vision (ECCV). (2016) 484–499
19. Wang, P.S., Liu, Y., Guo, Y.X., Sun, C.Y., Tong, X.: Adaptive o-cnn: A patch-based deep representation of 3d shapes. ACM Transactions on Graphics (TOG) **37** (2018) 1–11
20. Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3d object reconstruction from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017) 605–613
21. Insafutdinov, E., Dosovitskiy, A.: Unsupervised learning of shape and pose with differentiable point clouds. In: Advances in Neural Information Processing Systems (NeurIPS). (2018) 2802–2812
22. Kurenkov, A., Ji, J., Garg, A., Mehta, V., Gwak, J., Choy, C.B., Savarese, S.: Deformnet: Free-form deformation network for 3d shape reconstruction. In: Advances in Neural Information Processing Systems (NeurIPS). (2017) 858–866
23. Lin, C.H., Kong, C., Lucey, S.: Learning efficient point cloud generation for dense 3d object reconstruction. In: AAAI Conference on Artificial Intelligence (AAAI). (2018)
24. Wei, Y., Liu, S., Zhao, W., Lu, J., Zhou, J.: Conditional single-view shape generation for multi-view stereo reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019) 9651–9660
25. Groueix, T., Fisher, M., Kim, V.G., Russell, B., Aubry, M.: AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018) 216–224
26. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019) 4460–4470
27. Xu, Q., Wang, W., Ceylan, D., Mech, R., Neumann, U.: Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In: Advances in Neural Information Processing Systems (NeurIPS). (2019) 490–500
28. Cashman, T.J., Fitzgibbon, A.W.: What shape are dolphins? building 3d morphable models from 2d images. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) (2013) 232–244
29. Ntouskos, V., Sanzari, M., Cafaro, B., Nardi, F., Natola, F., Pirri, F., Ruiz, M.: Component-wise modeling of articulated objects. In: IEEE International Conference on Computer Vision (ICCV). (2015) 2327–2335
30. Kanazawa, A., Kovalsky, S., Basri, R., Jacobs, D.W.: Learning 3d deformation of animals from 2d images. In: Computer Graphics Forum. (2016) 365–374
31. Zuffi, S., Kanazawa, A., Jacobs, D., Black, M.: 3d menagerie: Modeling the 3d shape and pose of animals. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017) 6365–6373
32. Zuffi, S., Kanazawa, A., Black, M.J.: Lions and tigers and bears: Capturing non-rigid, 3D, articulated shape from images. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018) 3955–3963
33. Zuffi, S., Kanazawa, A., Berger-Wolf, T., Black, M.J.: Three-d safari: Learning to estimate zebra pose, shape, and texture from images "in the wild". In: IEEE International Conference on Computer Vision (ICCV). (2019) 5359–5368

34. Wen, C., Zhang, Y., Li, Z., Fu, Y.: Pixel2mesh++: Multi-view 3d mesh generation via deformation. In: IEEE International Conference on Computer Vision (ICCV). (2019) 1042–1051
35. Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N.: Semantic graph convolutional networks for 3d human pose regression. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019) 3425–3435
36. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016) 770–778
37. Kato, H., Ushiku, Y., Harada, T.: Neural 3d mesh renderer. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018) 3907–3916
38. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018) 586–595
39. Xiang, Y., Mottaghi, R., Savarese, S.: Beyond pascal: A benchmark for 3d object detection in the wild. In: IEEE Winter Conference on Applications of Computer Vision (WACV). (2014) 75–82
40. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2 (2019)
41. Kar, A., Tulsiani, S., Carreira, J., Malik, J.: Category-specific object reconstruction from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015) 1966–1974
42. Tulsiani, S., Zhou, T., Efros, A.A., Malik, J.: Multi-view supervision for single-view reconstruction via differentiable ray consistency. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017) 2626–2634