Learning-based Monocular 3D Reconstruction of Birds: A Contemporary Survey

Seyed Mojtaba Marvasti-Zadeh[*§], Mohammad N.S. Jahromi[†§], Javad Khaghani[x], Devin Goodsman[¶],
Nilanjan Ray[‡], and Nadir Erbilgin[*]
[*]Department of Renewable Resources, University of Alberta, Canada
[†]Research and Technology Department, Demant A/S, Copenhagen, Denmark
[x]Department of Electrical and Computer Engineering, University of Alberta, Canada
[¶]Canadian Forest Service, Natural Resources Canada, Edmonton, Alberta
[‡]Department of Computing Science, University of Alberta, Canada
Emails: {seyedmoj, khaghani, nray1, erbilgin}@ualberta.ca, mosj@demant.com, devin.goodsman@nrcan-rncan.gc.ca

***Abstract*—In nature, the collective behavior of animals, such as flying birds is dominated by the interactions between individuals of the same species. However, the study of such behavior among the bird species is a complex process that humans cannot perform using conventional visual observational techniques such as focal sampling in nature. For social animals such as birds, the mechanism of group formation can help ecologists understand the relationship between social cues and their visual characteristics over time (e.g., pose and shape). But, recovering the varying pose and shapes of flying birds is a highly challenging problem. A widely-adopted solution to tackle this bottleneck is to extract the pose and shape information from 2D image to 3D correspondence. Recent advances in 3D vision have led to a number of impressive works on the 3D shape and pose estimation, each with different pros and cons. To the best of our knowledge, this work is the first attempt to provide an overview of recent advances in 3D bird reconstruction based on monocular vision, give both computer vision and biology researchers an overview of existing approaches, and compare their characteristics.**

I. Introduction

Understanding the collective behavior of animals based on the interaction between individuals is an important task in many disciplines such as evolutionary biology, computational biology, and neuroscience [1], [2]. An excellent example of such behavior influencing social evolution can be observed for different bird species. For instance, changes in the visual pose and shape of flying birds over time help researchers understand and explain various aspects of social communication among birds through visual mechanisms [2]. In general, capturing such visual observation is performed by experts in the field over a continuous time period, often resulting in inaccurate measurements and observer bias. Consequently, to exploit the visual characteristics of birds, a reliable automated capture is required to estimate the birds' visual features such as pose and shape, respectively. However, estimating the pose and shape of birds in the wild is complicated for several reasons, such as the wide range of variations in shape and appearance [3].

Due to various technical challenges resulting from collecting multi-view 2D images or labeling 3D models in the wild, a new class of methods has emerged in 3D computer vision that attempts to reconstruct 3D objects from monocular vision (i.e., single-view images). Recent advances in 3D computer vision and deep learning have provided a variety of 3D reconstruction approaches from a 2D image collection or video sequence to estimate the pose and shape of a general object category. The proposed approaches are mainly benchmarked on two well-suited datasets, namely the Caltech-UCSD Birds (CUB 200-2007) [4] (6033 images of 200 bird species) and the extended version CUB 200-2011 [5] (11 788 images of 200 bird species) image collections tagged with keypoints, bounding boxes, coarse segmentation, and attribute labels. The main idea behind these approaches is to render 3D meshes from 2D image representation without the need to have synthetic data/multi-view images or 3D supervision [6]. However, the monocular 3D reconstruction of general object categories is an extremely challenging task due to the inherent ambiguities arising from various sources of geometric variability (e.g., different camera poses and shapes).

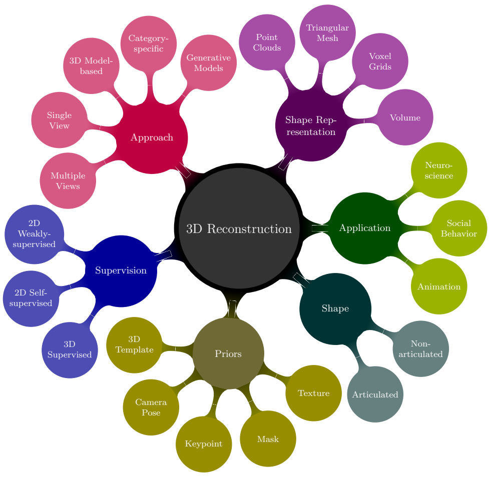

Fig. 1. Brief overview of 3D reconstruction methods for birds. This paper explores single-view, mesh-based approaches for articulated/non-articulated shapes utilizing various priors with 2D weak/self-supervisions.

The mesh-based approach is preferred among the existing approaches (e.g., generative-based [6] or model-based [7] (see Fig. 1)) because of its high shape expressivity (e.g., finer shape

[§] Equal contribution.

details). However, mesh-based methods depend on various types of supervision, such as initial 3D shapes, 2D semantic keypoints, and foreground masks. These priors or features extracted from 2D images serve as supervisory signals during learning when ground truth 3D shape is not available. To recover 3D structures from an image/video, these signals help the learning frameworks encode the 3D shape, camera pose, and texture associated with them.

Even though these learning-based methods have delivered promising results, tedious and careful 2D annotations are required that practically correspond to a handful of categories (i.e., limited generalization). Besides, the effectiveness of these methods can be dramatically affected by various issues such as large shape variations, invisible keypoints, and occlusion [8]. Furthermore, achieving impressive results requires acquiring large image collection and handling ambiguities associated with 2D observations across instances and viewpoints. Video sequences can be used as an alternative to image collection which is more convenient to deal with representing single/multi-view of the 3D shape of the same object. Moreover, utilizing videos can provide fine-grained information like optical flow to process motion beyond just instance keypoints or semantic parts [9]. There are still challenges to be overcome, such as reducing supervisory signals, providing generalization to other species, and dealing with data requirements and issues in real-world scenarios. To the best of our knowledge, no contribution has been made so far that has thoroughly reviewed the monocular 3D reconstruction of birds and its challenges from in-the-wild images/videos. Hence, this paper focuses on reviewing the learning-based image/video input approaches for 3D reconstruction of birds from 2D images and comparing their features, advantages, and limitations.

II. Monocular 3D Reconstruction: Taxonomy

In this section, the approaches for reconstructing 3D models of birds using monocular images/videos are discussed. Generally speaking, the objectives of these approaches are to reduce 2D supervision requirements and capture fine details through learning. Accordingly, the state-of-the-art approaches inferring the 3D structure of birds are compared in Table I.

A. Image Collection-based Approaches

A deeper understanding of 3D structure in images is contingent upon the global and local relationships between 2D percepts and 3D concepts. Hence, developing a computational model of birds can provide insight into their 3D structures. To learn this model, the well-known *category-specific mesh reconstruction* (CMR) [10] leverages 2D image collections annotated by foreground masks and semantic keypoint labels to recover 3D bird shape & texture and camera pose from a single image at inference. The CMR trains a ResNet-18-based encoder with three modules to efficiently predict deformable 3D mesh representations, assuming mirror-symmetry constraints and smooth surfaces. It explicitly learns to assign semantic keypoints of birds (such as legs & beaks) to a 3D mesh, which is enforced to be consistent with the foreground mask and natural world assumptions. In addition, the CMR relaxes the particular instance texturing to the fixed UV mapping of a mean shape with consistent semantic meaning. Although the CMR is able to recover overall shape (e.g., fat or thin birds), it cannot capture fine details of birds with asymmetrical articulation or rare poses.

As an extension of the CMR, the *unsupervised CMR* (U-CMR) [12] uses a single 3D template shape and a set of possible camera hypotheses (i.e., camera-multiplex) to address the requirement of keypoint labels and potential local minima problem for camera pose prediction. Inspired by particle filtering, the U-CMR utilizes a weakly-supervised iterative optimization approach to learn camera pose distribution for every shape and texture prediction. It comprises an encoder and two modules for predicting shape and texture, as well as threshold-based background subtraction and hole-filling to compute masks. The U-CMR can handle sharp long tails, protrusions of legs and beaks, and various bird types, although it fails to recover flying birds (large articulations) and birds with open wings or twisted heads due to its lack of an articulation model.

Besides, the *unsupervised mesh reconstruction* (UMR) [13] proposes a self-supervised model (to address the keypoint-dependency of the CMR [10]) that renders the semantic parts of each object to form a collection of deformable parts in both 2D and 3D space and their consistency (e.g., for wings on birds) across different instances of the similar category. It couples a collection of semantic parts (decomposed from a 2D image) with different object instances to build a canonical semantic UV texture map for each category, helping to create a template that captures major shape characteristics and semantic parts of the objects. Although the model is independent of any mesh or shape prior, the UMR involves some challenges such as poor segmentation method, rare camera pose instance and model limitation to capture certain shape characteristics details (e.g., the two wings of flying birds). For dealing with large shape variation and background distractors, the *weakly-supervised CMR* (WCMR) [9] utilizes a fusion module, shape-sensitive geometric constraints, and background manipulation to combine multi-scale features, supervise feature extraction & shape reconstruction, and improve robustness, respectively. By exploiting multi-scale features as well as the side-output mask, adaptive edge, & initial shape constraints, this method can cope with appearance variations, directly predict masks from the encoder's intermediate layers, be able to handle ambiguous shapes, and make explicit use of an initial shape.

On the other hand, the mapping of 2D image pixels to the locations on an abstract 3D model of a category can be considered the key to a rich understanding of objects and correspondence inference. Hence, the *canonical surface mapping* (CSM) [11] completes the cycle of pixels to 3D to pixels for birds by combining this mapping with 3D projection as well as exploiting a geometric cycle consistency loss to bypass reliance on keypoint or pose supervision. To do so, a parameterized network with U-Net architecture

TABLE I
COMPARISON OF LEARNING-BASED MONOCULAR 3D BIRDS RECONSTRUCTION.

			Input					Output				
	Method	Type of 2D Supervision	Keypoint	Mask	Camera pose	Template	Optical flow	3D Shape	Camera Pose	Texture	Deformation	Backbone
Image	CMR [10]	Weakly-supervised	✓	✓	SfM	initial shape	×	✓	✓	UV-flow	×	ResNet-18
	CSM [11]	Weakly-supervised	×	✓	×	✓	×	×	✓	×	×	ResNet-18
	U-CMR [12]	Weakly-supervised	×	✓	×	✓	×	✓	✓	✓	×	ResNet-18
	UMR [13]	Self-supervised	×	✓	×	×	×	✓	✓	UV-flow	×	ResNet-18
	WCMR [9]	Weakly-supervised	✓	✓	×	initial shape	×	✓	✓	UV-flow	✓	ResNet-18
	ACSM [14]	Weakly- or Self-supervised	×	✓	×	✓	×	×	✓	×	✓	ResNet-18
	IMR [8]	Weakly- or Self-supervised	×	✓	×	initial shape	×	✓	✓	UV-flow	×	ResNet-18
	TTP [15]	Weakly- or Self-supervised	×	✓	×	initial shape	×	×	✓	✓	✓	ResNet-18
	ACMR [16]	Weakly-supervised	✓	✓	×	×	×	✓	✓	UV-flow	✓	ResNet-18
	AVES [17]	Supervised	✓	✓	✓	articulated shape	×	✓	✓	×	✓	ResNet-50
Video	ACMR-vid [16]	Self-supervised	×	✓	×	×	×	✓	✓	UV-flow	✓	ResNet-50
	ACM [18]	Weakly-supervised	✓	✓	×	✓	✓	✓	✓	✓	✓	ResNet-18
	LASR [3]	Self-supervised	×	✓	×	×	✓	✓	✓	✓	✓	ResNet-18
	DOVE [19]	Self-supervised	×	✓	×	×	✓	✓	✓	✓	✓	Customized CNN

learns to infer canonical surface mapping using an input image and foreground mask with the aid of an additional network for predicting camera parameters. While an additional per-pixel mask predictor generates the foreground mask and ignores background pixels for inferring, a weak perspective transformation and a *neural mesh renderer* (NMR) [20] are employed to model camera projection and render a depth map, respectively. Despite the promising results, the primary limitations of the CSM approach pertain to handling significant shape differences or large articulations of categories (e.g., for non-rigid objects). As a solution to the CSM limitations, geometric consistency can be enforced without direct supervision to explicitly learn pose and articulation using input images and associated foreground mask labels. The *articulated CSM* (ACSM) [14] exploits this consistency and auxiliary losses to synchronize the pose, articulation, and canonical surface mapping and prevent degenerate or trivial solutions, respectively. It employs two ResNet-18-based encoder-decoder networks to jointly infer the mapping as well as the camera pose and articulation of 3D templates, while it can be extended to scenarios with available annotations. However, this method is restricted to intrinsic shape variations of species and results in sub-optimal performance for birds due to its ambiguities for their parts (e.g., beak and wing).

Despite the approaches that separate modules of a network regress camera pose and non-rigid deformation, the *To-The-Point* [15] (TTP) jointly learns them by a lightweight per-sample differentiable optimization given 2D images, foreground masks, and optional keypoints. This method uses a sampling-based texture approach and as-rigid-as-possible (ARAP) constraint [21] to estimate accurate correspondence predictions and constraint arbitrary deformations (e.g., anomalies). The TTP uses Mask-RCNN and ResNet-based encoder to obtain foreground masks and map an image to latent feature maps, respectively. Although the mesh can be efficiently recovered in a few iterations (or even one), inaccurate camera poses and incorrect 2D points can negatively impact its performance (e.g., for flying birds).

The learning of explicit 3D representations entails limitations regarding scalability beyond a few categories as well as coping with intra-instance shape variations such as thin and fat birds. However, inferring the implicit shape and texture of a 3D object generally relies on 3D supervision and templates. The *implicit mesh reconstruction* (IMR) [8] addresses these restrictions through learning 3D inference with only in-the-wild image collections along with rough object segmentation. The IMR uses geometric consistency to bypass the need for direct supervision, thereby promoting implicit shape and texture representations. After obtaining an initial 3D mesh from an encoder, this method parameterizes the shape and texture to learn the category-level latent representations. Besides, it uses a simple four-layer feedforward MLP to instantiate the shape and texture implicit functions. Although this method addresses several bottlenecks and provides good prediction of shape and texture for different object categories, it is limited to the rendering of objects that are not subject to large shape variations. In addition, the IMR assumes that the object is not occluded, which may affect the overall performance.

Despite relying on weakly- or self-supervised approaches, trained networks may predict unstable and inconsistent shapes for video sequences because of perturbations over time. The *asymmetric CMR* (ACMR) trains an image-based network on a collection of category-specific images with an ARAP constraint [21] to predict the shape, texture, and camera pose and avoid the symmetric assumption to effectively handle scenarios such as birds tilting/rotating their heads or deforming from standing to flying. This method uses a weighted combination of learned shape bases to recover asymmetric shapes and prevent large motion deformations of shape models, which can be trained by an encoder using either object silhouettes & 2D keypoints or only object silhouettes. The ACMR also can exploit online adaptation to generalize the learned model to input videos and self-supervised setting. For online adaptation, all parameters of the learned reconstruction model are fine-tuned, and then random parts onto the middle frame are painted and propagated backward to the first frame and forward to the last frame. During self-supervised training, this method stabilizes the predictions by taking shape symmetry into account, as opposed to its weakly-supervised setting that benefits from keypoint supervision. Lastly, to deform the generic template model to new species and the individual instances sequentially, the *Avian species-specific* (AVES) [17] extends the articulated bird mesh obtained in [2] considering lack of strong deformable shape prior. It aligns a generic template model with the samples and then deforms it to new species to acquire the mean shape of them. The mean shape is further deformed to capture variations among different instances of the new species using a combination of shape basis.

B. Video Frames-based Approaches

The dense correspondences between consecutive frames can be leveraged as the self-supervision to reconstruct the 3D shapes of articulated objects. The *articulated categories from motion* (ACM) [18] uses this supervision to optimize regularized mesh deformations using per-frame semantic segmentation masks and keypoint reprojections. This method exploits single-frame networks (i.e., one ResNet-based encoder and three decoders for predictions) that jointly infer category objects' 3D shape, camera pose, and texture. It can correctly capture the wing variation of birds (highly flexible wings), which is challenging for alternative methods. However, incorrect global camera parameters (e.g., scale) can cause failures caused by the small diversity of appearances in videos.

To avoid relying on category-specific 3D shape templates, the *learning articulated shape reconstruction* (LASR) [3] jointly learns 3D shape, articulation, and camera parameters from a single monocular video to provide a template-free approach. The joint recovery mechanism is obtained by solving inverse graphics problem via gradient-based optimization. It employs a network to parameterize the camera parameters and articulation. Because LASR relies on generic shape and motion priors, it can be applied to a wider range of non-rigid shapes. However, the main drawback of LASR is failure to render invisible surface of occluded objects that are missed by mask annotations.

Finally, the *deformable objects from videos* (DOVE) [19] method employs video frames from short clips, corresponding 2D masks, and optical flows to reconstruct an articulated 3D shape with texture and rigid pose of an object instance, without requiring category-specific 3D template. Adopting a photo-geometric auto-encoding, this method takes a base mesh to optimize its vertices by three networks (rigid pose, shape, & texture), whose recombined information provides supervision. Despite the simple geometry and deformation structure of birds, the performance of this method can be degraded when the birds are away from the camera or have pose variations of more than 90 degrees. In addition, the results may be affected by the required segmentation masks, motion blur, or missing fine details (e.g., legs and beak).

III. Conclusion

In this paper, we presented recent advances to the learning-based 3D reconstruction of bird species using single-view images and videos. The challenging idea of recovering the pose and shape of birds, which can play an important role in understanding the collective behavior of species in the wild, has inspired a new research direction in the area of computational ecology and neuroscience. Although each approach in this area shows superior results under certain challenging scenarios, learning schemes, and geometric variations, we acknowledge that there are still several challenges that exist to solving this highly ill-posed problem. However, we believe that highlighting the strengths and weaknesses of each 3D reconstruction method for flying birds provides great value for future studies in this research direction.

Acknowledgments

This work was supported in-part by the fRI Research-Mountain Pine Beetle Ecology Program.

References

[1] M. Vidal, N. Wolf, B. Rosenberg, B. P. Harris, and A. Mathis, "Perspectives on individual animal identification from biology and computer vision," *Integrative and comparative biology*, vol. 61, no. 3, pp. 900–916, 2021.

[2] M. Badger, Y. Wang, A. Modh, A. Perkes, N. Kolotouros, B. G. Pfrommer, M. F. Schmidt, and K. Daniilidis, "3D bird reconstruction: A dataset, model, and shape recovery from a single view," in *Proc. ECCV*, 2020, pp. 1–17.

[3] G. Yang, D. Sun, V. Jampani, D. Vlasic, F. Cole, H. Chang, D. Ramanan, W. T. Freeman, and C. Liu, "LASR: Learning articulated shape reconstruction from a monocular video," in *Proc. CVPR*, 2021, pp. 15 980–15 989.

[4] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, "Caltech-UCSD birds 200," *Computation & Neural Systems Technical Report, California Institute of Technology*, 2010.

[5] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-USCD birds-200-2011 dataset," *Computation & Neural Systems Technical Report, California Institute of Technology*, 2011.

[6] D. Pavllo, J. Kohler, T. Hofmann, and A. Lucchi, "Learning generative models of textured 3D meshes from real-world images," in *Proc. ICCV*, 2021, pp. 13 879–13 889.

[7] B. Biggs, O. Boyne, J. Charles, A. Fitzgibbon, and R. Cipolla, "Who left the dogs out? 3D animal reconstruction with expectation maximization in the loop," in *Proc. ECCV*, 2020, pp. 195–211.

[8] S. Tulsiani, N. Kulkarni, and A. Gupta, "Implicit mesh reconstruction from unannotated image collections," 2020. [Online]. Available: https://arxiv.org/abs/2007.08504

[9] S. Sun, Z. Zhu, X. Dai, Q. Zhao, and J. Li, "Weakly-supervised reconstruction of 3D objects with large shape variation from single in-the-wild images," in *Proc. ACCV*, 2020.

[10] A. Kanazawa, S. Tulsiani, A. A. Efros, and J. Malik, "Learning category-specific mesh reconstruction from image collections," in *Proc. ECCV*, 2018, pp. 371–386.

[11] N. Kulkarni, A. Gupta, and S. Tulsiani, "Canonical surface mapping via geometric cycle consistency," in *Proc. ICCV*, 2019.

[12] S. Goel, A. Kanazawa, and J. Malik, "Shape and viewpoint without keypoints," in *Proc. ECCV*, 2020, pp. 88–104.

[13] X. Li, S. Liu, K. Kim, S. D. Mello, V. Jampani, M.-H. Yang, and J. Kautz, "Self-supervised single-view 3D reconstruction via semantic consistency," in *Proc. ECCV*, 2020, pp. 677–693.

[14] N. Kulkarni, A. Gupta, D. F. Fouhey, and S. Tulsiani, "Articulation-aware canonical surface mapping," in *Proc. CVPR*, 2020.

[15] F. Kokkinos and I. Kokkinos, "To the point: Correspondence-driven monocular 3D category reconstruction," in *Proc. NIPS*, vol. 34, 2021, pp. 7760–7772.

[16] X. Li, S. Liu, S. De Mello, K. Kim, X. Wang, M.-H. Yang, and J. Kautz, "Online adaptation for consistent mesh reconstruction in the wild," in *Proc. NIPS*, vol. 33, 2020, pp. 15 009–15 019.

[17] Y. Wang, N. Kolotouros, K. Daniilidis, and M. Badger, "Birds of a feather: Capturing avian shape models from images," in *Proc. CVPR*, 2021, pp. 14 739–14 749.

[18] F. Kokkinos and I. Kokkinos, "Learning monocular 3D reconstruction of articulated categories from motion," in *Proc. CVPR*, 2021, pp. 1737–1746.

[19] S. Wu, T. Jakab, C. Rupprecht, and A. Vedaldi, "DOVE: Learning deformable 3d objects by watching videos," 2021. [Online]. Available: https://arxiv.org/abs/2107.10844

[20] H. Kato, Y. Ushiku, and T. Harada, "Neural 3D mesh renderer," in *Proc. CVPR*, 2018.

[21] O. Sorkine and M. Alexa, "As-rigid-as-possible surface modeling," in *Geometry Processing*, 2007.