Online Adaptation for Consistent Mesh Reconstruction in the Wild

Xueting Li[1], **Sifei Liu**[2], **Shalini De Mello**[2], **Kihwan Kim**[2], **Xiaolong Wang**[3], **Ming-Hsuan Yang**[1], **Jan Kautz**[2]
[1]University of California, Merced, [2]NVIDIA, [3]University of California, San Diego

Abstract

This paper presents an algorithm to reconstruct temporally consistent 3D meshes of deformable object instances from videos in the wild. Without requiring annotations of 3D mesh, 2D keypoints, or camera pose for each video frame, we pose video-based reconstruction as a self-supervised online adaptation problem applied to any incoming test video. We first learn a category-specific 3D reconstruction model from a collection of single-view images of the same category that jointly predicts the shape, texture, and camera pose of an image. Then, at inference time, we adapt the model to a test video over time using self-supervised regularization terms that exploit temporal consistency of an object instance to enforce that all reconstructed meshes share a common texture map, a base shape, as well as parts. We demonstrate that our algorithm recovers temporally consistent and reliable 3D structures from videos of non-rigid objects including those of animals captured in the wild – an extremely challenging task rarely addressed before. Codes and other resources will be released at https://sites.google.com/nvidia.com/vmr-2020.

1 Introduction

When we humans try to understand the object shown in Fig. 1(a), we instantly recognize it as a "duck". We also instantly perceive and imagine its shape in the 3D world, its viewpoint, and its appearance from other views. Furthermore, when we see it in a video, its 3D structure and deformation become even more apparent to us. Our ability to perceive the 3D structure of objects contributes vitally to our rich understanding of them.

While 3D perception is easy for humans, 3D reconstruction of deformable objects remains a very challenging problem in computer vision, especially for objects in the wild. For learning-based algorithms, the key bottleneck is the lack of supervision. It is extremely challenging to collect 3D annotations such as 3D shape and camera pose [4, 17]. Consequently, existing research mostly focuses on limited domains (e.g., rigid objects [24], human bodies [14, 50] and faces [47]) for which 3D annotations can be captured in constrained environments. However, these approaches do not generalize well to non-rigid objects captured in naturalistic environments (e.g., animals). In non-rigid structure from motion methods [2, 28], the 3D structure can be partially recovered from correspondences between multiple viewpoints, which are also hard to label. Due to constrained environments and limited annotations, it is nearly impossible to generalize these approaches to the 3D reconstruction of non-rigid objects (e.g., animals) from images and videos captured in the wild.

Instead of relying on 3D supervision, weakly supervised or self-supervised approaches have been proposed for 3D mesh reconstruction. They use annotated 2D object keypoints [13], category-level templates [21, 20] or silhouettes [22]. However, to scale up learning with 2D annotations to hundreds of thousands of images is still non-trivial. This limits the generalization ability of current models to new domains. For example, a 3D reconstruction model trained on single-view images, e.g., [13], produces unstable and erratic predictions for video data. This is unsurprising, due to perturbations over time. However, the temporal signal in videos should provide us an advantage instead of a

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

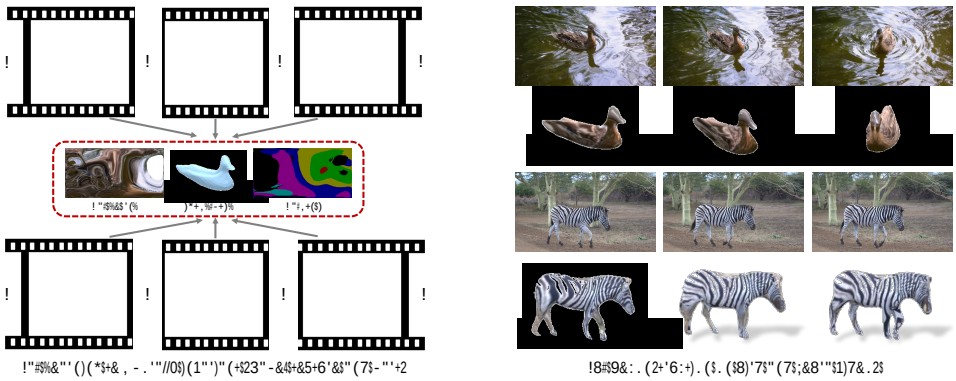

Figure 1: By utilizing the consistency of texture, shape and object parts correspondences in videos (red box) as self-supervision signals in (a), we learn a model that reconstructs temporally consistent meshes of deformable object instances in videos in (b).

disadvantage, as recently shown on the task of optimizing a 3D *rigid* object mesh w.r.t. a particular video [53, 24]. The question is, can we also take advantage of the redundancy in temporal sequences as a form of self-supervision in order to improve the reconstruction of dynamic non-rigid objects?

In this work, we address this problem with two important innovations. First, we strike a balance between model generalization and specialization. That is, we train an image-based network on a set of images, while at test time we adapt it online to an input video of a particular instance. Test-time training [38] is non-trivial since no labels are provided for the video. The key is to introduce self-supervised objectives that can continuously improve the model. To do so, we exploit the UV texture space, which provides a parameterization that is invariant to object deformation. We encourage the sampled texture, as well as a group of object parts, to be consistent among all the individual frames in the UV space, as shown in Fig. 1(a). Using this constraint of temporal consistency, the recovered shape and camera pose are stabilized considerably and are adapted to the current video.

One bottleneck of existing image-based 3D mesh reconstruction methods [13, 22] is that the predicted shapes are assumed to be symmetric. This assumption does not hold for most non-rigid animals, e.g., birds tilting their heads, or walking horses, etc. Our second innovation is to remove this assumption and to allow the reconstructed meshes to fit more complex, non-rigid poses via an as-rigid-as-possible (ARAP) constraint. As another constraint that does not require any labels, we enforce ARAP during test-time training as well, to substantially improve shape prediction. We use two image-based 3D reconstruction models for training (i) a weakly supervised one (i.e., with object silhouettes and 2D keypoints provided), and (ii) a self-supervised one where only object silhouettes are available. The image-based models are then adapted to in-the-wild bird and zebra videos collected from the internet. We show that for both models, our innovations lead to an effective and robust approach to deformable, dynamic 3D object reconstruction of non-rigid objects captured in the wild.

2 Related Work

3D object reconstruction from images. A triangular mesh has long been used for object reconstruction [13, 17, 25, 15, 44, 29, 45]. It is a memory-efficient representation with vertices and faces, and is amenable to differentiable rendering techniques [17, 25]. The task of 3D reconstruction entails the simultaneous recovery of the 3D shape, texture, and camera pose of objects from 2D images. It is highly ill-posed due to the inherent ambiguity of correctly estimating both the shape and camera pose together. A major trend of recent works is to gradually reduce supervision from 3D vertices [4, 45, 44], shading [10], or multi-view images [49, 17, 46, 33, 24] and move towards weakly supervised methods that instead use 2D semantic keypoints [13], or a category-level 3D template [21]. This progress makes the reconstruction of objects, e.g., birds, captured in the wild possible. More recently, self-supervised methods [22, 47, 16] have been developed to further remove the need for annotations. Our method exploits different levels of supervision: weak supervision (i.e., using 2D semantic keypoints) and self-supervision to learn an image-based 3D reconstruction network from a collection of images of a category (Sec. 3.1).