We thank the reviewers for their valuable comments and suggestions. We address the common issues before diving into the detailed questions for each reviewer. We will cite the reference [7] mentioned by R4 in the revised paper.

[R1, R2, R4] Significance and Scope. Our work focuses on mesh reconstruction of animals *in the wild* for which 3D scans or parametric models cannot be easily obtained. Our problem setting aligns with previous closely-related work on single-view mesh reconstruction – CMR [13], UMR [21] and UCMR (released after NeurIPS submission). More importantly, we contribute to this line of research by exploring reconstruction of dynamic objects from unlabeled videos, which NONE of these previous works has explored. We show how our approach can help to improve 3D reconstruction within the context of this very challenging problem by exploiting asymmetric reconstruction and temporal coherence.

[R1, R2, R4] Generalizability. The reviewers request us to evaluate on humans (R2, R4), monkeys (R2), faces (R4), and quadrupeds (R4), for which strong parametric 3D prior models are manually designed by scanning numerous 3D instances (or figurines) carefully in controlled environments. Our method in contrast, is designed for unconstrained settings without strong shape priors or annotations. We assume a simple prior, e.g., a sphere, without knowing articulations or parts. Even though we model the shapes via shape bases, our bases are learned in a data-driven manner by clustering estimated shapes. We follow related works CMR [13] and UMR [21], which are all validated primarily on the bird category. Furthermore, we also test on a quadruped category (zebras). Since our method deals with a much harder problem with more shape/camera ambiguity, it is less relevant for categories with existing strong shape priors.

[R2] Shape Variations. Our reconstructed dynamic meshes do show large, asymmetric shape variations, e.g., bird rotating head (Fig. 5, paper), flying (Fig. 4,6, supp; Fig. 1), and zebra walking (Fig. 9, supp; demo). Note that shape variation to this degree has NOT been tackled before by any of the close-related SOTA works, including CMR (Fig. 6 in the supp). Meanwhile, existing methods that tackle large shape variation [54,55] achieve it by using much stronger parametric models, e.g. SMAL [54] *v.s.* our basis shape model, which is learned via weaker supervision.

[R3] Quantitative Evaluation. We note that the keypoint transfer metric is effective in evaluating camera and texture prediction [13,16,20]. To further quantitatively evaluate shape reconstruction quality, we animate a synthetic 3D bird model and create a video with 520 frames in various poses such as flying, landing, walking etc., as shown in Fig. 1 below. We compare predicted meshes with ground truth meshes using Chamfer distance every 10 frames and show the evaluation results in Table 2 below. The proposed ACMR method outperforms the CMR [13] model and is further improved via the proposed online adaptation strategy. More videos will be created and evaluated in our revised paper.

[R3] Clarity. We use weak-perspective projection like CMR. We use 8 blend shapes (Fig.10(a), supp). The full ACMR model is in Table 1(c), not Table 1(f) in the paper. Our unsupervised model is modified from ACMR (Sec 3.3), where the only part based on UMR is for template learning (Sec 3.3 (ii)). We present the unsupervised method to show the generalization of our online adaptation. We will include more details on this part to make it reproducible.

[R4] Novelty. In comparison to references [3-6] noted by R4, our task is quite different and more difficult in terms of shape priors and video labels, etc (Table 1 below). Without parametric shape priors, our consistency module needs to be more powerful to correct the camera/shape confusion. All the other works [3-6] do not need to explicitly deal with this. We further note that a fair comparison against [3-6], would require re-implementing these methods in their entirety. Without a comparable setting, it is not meaningful to compare individual blocks of the various algorithms in isolation. We also note that we didn't claim addressing or mention "blurry texture" or "low-res mesh" in the paper.

[R4] Static Shape. The key to upgrading image-based methods to videos is to model shape asymmetry, so that animals can move (line 129, paper) – this has never been explored in any of the close-related static methods, e.g., CMR or UMR (CSM [20] cannot perform reconstruction). We extensively validate the contribution of our static model qualitatively (Fig. 4, paper; Fig. 4,6, supp), and quantitatively (Tab. 1, paper; Tab. 1,2, supp). The key idea of keypoint re-projection is *category-level* integration (Fig. 3(c)) of individual instances in order to better leverage their semantic invariance in the UV space. In contrast, CSM's cycle consistency is conducted at the *instance-level* and never explores such invariance.

[R4] Existing work with videos [5,6]. Both [5,6] study the human body based on SMPL, which largely reduces the difficulty. They are also provided with labeled 3D, or a combination of 2D and 3D ground truth on videos (Table 1 below). Thus their temporal consistency is largely guaranteed via supervised learning. Additionally, neither models textures and appearance at all. Hence, they are significantly different from and not directly comparable to our work.

[R4] Parametric Model & Deformation. The SMPL model is learned using ground truth 3D annotations, while our bases shapes are not. The two are not directly comparable. Regarding the degree of deformation, since the base shape is a blended version and L_s are computed via a sliding local window, our method is able to handle base shape transitions over time. One example is shown in Fig. 1 below, the second row.

Ref.	parametric model	any video label?	controlled	consistency
[3]	DAM	ground truth camera	Yes	texture
[4]	kinematics	ground truth camera	Yes	keypoints
[5]	SMPL	2D & 3D ground truth	Yes	shape
[6]	SMPL	3D ground truth	Yes	supervised
ours	none	one unlabeled video	No	texture, shape, parts

Table 1: Comparisons of settings with [3-6] in R4's review.

Figure 1: Reconstruction of animated synthetic 3D bird video.

Method	Chamfer↓
CMR[13]	0.016
ACMR	0.015
ACMR (T)	**0.012**

Table 2: Evaluation on synthetic 3D bird videos.