Reconstructing Animatable 3D Categories from Videos:
SUPPLEMENTARY MATERIAL

A. Additional Details

Shape Regularization. We apply eikonal regularization to force the norm of the first order derivative of signed distances d to be close to 1,

$$\mathcal{L}_{\text{eikonal}} = (\|\nabla \mathbf{MLP}_{\text{SDF}}(\mathbf{X})\| - 1)^2. \quad (1)$$

Eikonal loss forces the reconstruction to be a valid surface and empirically improves the surface reconstruction quality.
Pose, Deformation, and Appearance Smoothness. We would like the time-varying articulated pose, deformation, and appearance codes $\{\boldsymbol{\theta}, \omega_d, \omega_a\}$ to vary smoothly within a video. To accomplish this, we make use of time-dependent positional embeddings (similar to [11]):

$$\omega_t^b = \mathbf{A}_i \mathcal{F}(t) \quad (2)$$

where $\mathcal{F}(\cdot)$ is a 1D basis of sines and cosines with linearly-increasing frequencies at log-scale [6], and we learn separate weight matrices $\mathbf{A}_{i\in\{1\dots,M\}}$ for each video.

B. Category Outside DensePose

We test RAC in a scenario where there is no predefined DensePose features and skeleton.

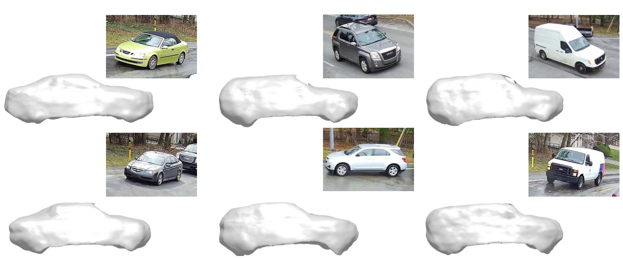

Figure 1. **Vehicle Category Reconstruction.** Our method is able to fuse videos of 365 vehicles with different appearance and shape into a category model. From left to right, we show reconstruction of sedans, SUVs, and vans.

Vehicle Dataset. We employ images from multiple 4K cameras [4] that overlook urban public spaces to analyze the flow of traffic vehicles. The data are captured for 3-second bursts every few minutes, and only images with notable changes are stored. We extracted 365 car videos from the dataset to build the category model. The dataset contains wide variation in vehicle categories like pickup trucks, construction vehicles etc on which traditional model based approached perform poorly.
Camera Pose Initialization. As there is no DensePose model for cars, we took a two-stage approach to first coarsely register a few car videos with manual viewpoint annotation and then train a single-image viewpoint network to predict viewpoints for the rest of the videos. The camera viewpoints are roughly annotated for each frame (with around 30 degree rotation error). Annotation for a 100 frame video takes around 30 seconds. We found annotating 20 cars to be sufficient to train a viewpoint estimator that generalizes to other cars.
Results. We show the reconstruction results of car videos in Fig. 1. Please visit the website for more results.

Table 1. **Quantitative results on `Pablo` sequence.** 3D Chamfer distance (cm, ↓) is computed on the clothing region and averaged over all frames. MPCap uses a pre-scanned personalized template.

Method	MPCap*	MCCap	PiFuHD	T2S	
Chamfer	14.6	17.9	26.5	27.7	18.3

C. Evaluation on `Pablo` Sequence

We compare with baselines on the `Pablo` sequence, which is part of the public MonoPerfCap [10] dataset. Our method optimizes the `Pablo` sequence together with the rest of our 47 human videos. After differentiable rendering optimization, we extract meshes for the `Pablo` sequence and compare with the 3D ground-truth for evaluation.
Metrics. We follow the evaluation protocol of MonoClothCap [9] and compute the average point-to-surface distances in the clothing region. The clothing region (the T-shirt and shorts) is obtained by manual segmentation on the ground-truth surface mesh.
Results. We show quantitative comparisons in Tab. 1 and refer the reader to the qualitative results in Fig. 8 of the main draft. Our method outperforms PiFuHD [5], Tex2Shape (T2S) [1], both of which are single-view human shape pre-

dictors trained on 3D scans of humans. Our method does not use 3D data to train but performs test-time optimization on 47 human videos. Our method is slightly worse than MonoClothCap (MCCap) [9] that uses a parametric human body model (SMPL), and worse than MonoPerfCap (MP-Cap), which uses a prescanned template. Both parametric body model and personalized shape template provides a strong shape prior, while our method does not rely on any shape prior.

D. Difference from prior works

We highlight the difference from previous work in Tab. 2. In terms of shape modeling, HyperNeRF [3] and HumanNeRF [7] reconstruct a *single* scene or instance, while learns a space of category shapes. For skeleton modeling, CASA [8] is optimized *per-instance*, while learns a shared space over a category of skeletons (with different bone lengths). For background modeling, NeRF++ [12] assumes a static scene and does not use background to help object segmentation and reconstruction. NerFace [2] treats background as a static image, while we represent the background as a NeRF, which generalizes to videos captured by a moving camera.

Table 2. **Difference between prior works and .**

Method	Shape	Motion	Background	3D Data/Pose
NeRF++	N.A.	N.A.	NeRF	No
NeRFace	Instance	Conditional	Image	No
HyperNeRF	Instance	Fields+Conditional	N.A.	No
BANMo	Instance	Control Points	N.A.	No
CASA	Instance	Instance Skeleton	N.A.	Yes
HumanNeRF	Instance	Instance Skeleton	N.A.	Yes
	Category	Category Skeleton	NeRF	No

References

[1] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *ICCV*, pages 2293–2303, 2019. 1

[2] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021. 2

[3] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv*, 2021. 2

[4] N. Dinesh Reddy, Robert Tamburo, and Srinivasa G. Narasimhan. Walt: Watch and learn 2d amodal representation from time-lapse imagery. In *CVPR*, pages 9356–9366, June 2022. 1

[5] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, 2020. 1

[6] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *NeurIPS*, 2020. 1

[7] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *CVPR*, pages 16210–16220, June 2022. 2

[8] Yuefan Wu, Zeyuan Chen, Shaowei Liu Liu, Zhongzheng Ren, and Shenlong Wang. Casa: Category-agnostic skeletal animal reconstruction. In *NeurIPS*, 2022. 2

[9] Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica Hodgins. Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video. In *3DV*, pages 322–332. IEEE, 2020. 1, 2

[10] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Trans. Graph.*, 37(2):27:1–27:15, May 2018. 1

[11] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *CVPR*, 2022. 1

[12] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 2