



End-to-End Learning of Multi-category 3D Pose and Shape Estimation

Yigit Baran Can¹, Alexander Liniger¹, Danda Pani Paudel¹, Luc Van Gool^{1,2}
¹Computer Vision Lab. ETH Zurich: ²VISICS, KU Leuven

1 Introduction

A keypoint-based shape and pose representation is attractive because of its simplicity and ease of handling. In this paper, we consider that only one image per object is available in both during training and inference. We also assume that only minimalistic supervision in the form of 2D keypoints and objects' categories are available. Proposed method is an end-to-end multi class deep non-rigid structure-from-motion (NrSfM) that only takes an image as input and outputs canonical 3D shape and camera pose.

2 Related Work

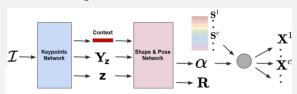
C3DPO[1] learns the factorization of the object deformation and viewpoint change. Transversal property through a separate canonicalization network. Procrustean regression is used to determine unique motions and shapes[2]. Also an end-to-end method using a CNN that can output 3D location of human keypoints from the image. Yet, it cannot handle multiple object categories or occluded keypoints. Human pose estimation is also tackled in[3], where the authors propose a cyclic-loss and discriminator. Recently [4] extended Procrustean formulation with autoencoders and proposed a method that can infer 3D shapes. Most methods accept 2D keypoints as input rather than images and tackle the problem of obtaining 3D keypoint locations from a single image using a separate keypoint detector, such as a stacked hourglass network [5].

3 Method

• The method is in the framework of Non-Rigid Structure from Motion (NrSfM). Specifically, the 3D keypoint locations are represented as linear combination of a dictionary $\mathbf S$. The matrix $\mathbf S$ is formed from the basis vectors where the coefficients for the basis vectors are encoded by $\mathbf \alpha$. The camera rotation matrix R is estimated together with $\mathbf S$ and $\mathbf \alpha$. π represents the projection matrix.

$$\min_{\alpha_i, \mathbf{S}, \mathbf{R}_i \in \mathbf{SO(3)}} \quad \sum_{i=1}^n \mathcal{L}(\mathbf{Y}_i, \Pi \mathbf{R}_i(\mathbb{I}_3 \odot \alpha_i^{\mathsf{T}} \mathbf{S})).$$

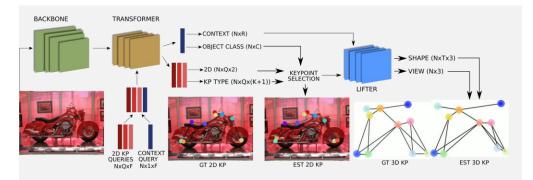
 The proposed method takes an image as input and estimates the 2D keypoints as an intermediate step as well as producing a context vector to guide the 3D estimation.



The method is end-to-end but operationally can be examined in 2 parts: 2D-3D (lifter) network and the 2D estimation network. For lifter, we propose cut-off coefficients. Applying ReLU on α results in sparse selection of basis vectors. To gain back the expressiveness, we introduce the bias term b_s.

$$\mathcal{L}(\mathbf{Y} \circ \zeta, \Pi(R(\mathbf{Y})(\text{ReLU}(W_q f(\mathbf{Y}) + b_q)\mathbf{S} + b_S) \circ \zeta))$$

 The 2D network is a transformer where the query vectors produce the 2D locations and the semantics of the keypoints. The estimated keypoints are then fed to the lifter network. A separate query vector produces contextual information and directly connects to the lifter network to provide a skip connection to the transformer layers from the end task.



The total loss function is in terms of 2D estimates Y, semantics w and object class z: Location loss $\mathcal{L}_l = ||\bar{\mathbf{Y}} - \hat{\delta}||_1$

Category loss $\mathcal{L}_b = \mathcal{L}_{CE}(\bar{\mathbf{z}}, \hat{\mathbf{z}})$ KP Type loss $\mathcal{L}_k = \mathcal{L}_{CE}(\Omega, \hat{\omega})$

Reprojection loss $\mathcal{L}_r = \mathcal{L}(\bar{\mathbf{Y}} \circ \zeta, \Pi R(\hat{\mathbf{Y}}, \rho) f(\hat{\mathbf{Y}}, \rho) \circ \zeta)$

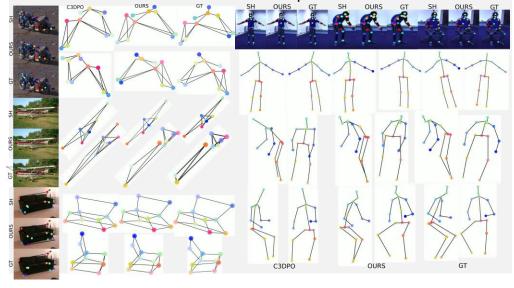
4 Results

We report results when our method uses GT 2D keypoints as input (Left) and the results for directly image-to-3D (Right) .

	Pascal3D		Human3.6M		S-Up3D	
Method	MPJPE	Stress	MPJPE	Stress	MPJPE	Stress
Geo-SH ‡*	-	-	51	-	-	-
EM-SfM	131.0	116.8	-	-	0.107	0.061
GbNrSfM	184.6	111.3	-	-	0.093	0.062
PoseGAN	-	-	130.9	51.8	-	-
Proc †*	-	-	86.4	-	-	-
PAUL †	30.9	-	88.3	-	0.058	-
C3DPO-base	53.5	46.8	135.2	56.9	0.160	0.105
C3DPO	38.0	32.6	101.8	43.5	0.068	0.040
Ours	29.5	26.6	92.8	42.6	0.057	0.035

	I docuis D		1101110110110111		
Method	MPJPE	Stress	MPJPE	Stress	
Geo-SH ‡*	-	-	68	-	
CMR/SH	74.4	53.7	-	-	
C3DPO/SH	57.4	41.4	145.0	84.7	
Proc-SH †*	-	-	124.5	-	
Proc-CNN †*	-	-	108.9	-	
PAUL-SH †	-	-	132.5	-	
Ours/SH	56.1	39.0	140.7	80.9	
Ours/TR	61.3	47.9	114.0	58.8	
Ours w/o Cont	57.6	42.9	113.8	56.7	
Ours	51.6	35.3	107.7	55.4	

The proposed cut-off formulation produces huge improvements and the overall end-to-end method surpasses the SOTA.



5 Conclusion

We study estimating 3D pose and shape from a single image for objects of multiple categories, in an end-to-end manner by only using 2D keypoint annotations for supervision. Results show that end-to-end training and the use of contextual information improve the performance substantially. Our method is the first of its kind, providing a framework that can be applied to any dataset. We outperform all the compared methods on three datasets.

References

- 1. Nov otny et al. C3DPO: Canonical 3D Pose Networks for Non-Rigid Structure From Motion. ICCV 2019.
- Park et al. Procrustean Regression Networks: Learning 3D Structure of Non-rigid Objects from 2D Annotations. ECCV 2020..
- 3. Chen et al. Unsupervised 3D Pose Estimation With Geometric Self -Supervision. CVPR 2019.
- Wang et al. PAUL: Procrustean Autoencoder for Unsupervised Lifting. CVPR 2021.
- 5. Toshev et al. DeepPose: Human Pose Estimation via Deep Neural Networks. CVPR 2014.