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

Yigit Baran Can¹ yigit.can@vision.ee.ethz.ch Alexander Liniger¹ alex.liniger@vision.ee.ethz.ch Danda Pani Paudel¹ paudel@vision.ee.ethz.ch Luc Van Gool^{1,2} vangool@vision.ee.ethz.ch ¹ Computer Vision Lab ETH Zurich ² VISICS, ESAT/PSI KU Leuven

1 Expressiveness Proof

Now let us show that this formulation does not affect the expressiveness of the network, even if we would not re-optimize the basis **S**. Let a sample Y_n , be expressed with the basis coefficients $\alpha_n \in \mathbb{R}^D$, i.e. $\psi_n = \alpha_n \mathbf{S}$. Let the proposed representation with cut-off coefficients be $\hat{\psi}_n = \beta_n \mathbf{S} + b_s$. Let us index the dimensions of *D* dimensional latent space with *d*. Then the latent space coefficient of sample Y_n for the dimension *d* is α_{nd} , and similarly for $\hat{\psi}_n$, it is β_{nd} . The question is, can we find β_n and b_s such that $\hat{\psi}_n = \psi_n$ for all *n*. If this holds, we do not lose any expressive power while arriving in a differentiable manifold selection rule. In mathematical terms this is fulfilled if $\exists \beta_{nd}, b_s$ such that $\sum_d \alpha_{nd} \mathbf{S}_d = \sum_d \beta_{nd} \mathbf{S}_d + b_s$, $\forall n \leq N$. Note that the bias, which is independent of *n*, is fundamental.

Let $m_d = \min_n \alpha_{nd}$ or in other terms, m_d is the minimum coefficient for latent dimension d among all the data samples $n \leq N$. Then, $\alpha_{nd} - m_d \geq 0, \forall n, d$. This implies that we can set $\beta_{nd} = \alpha_{nd} - m_d$ and $b_S = \sum_d m_d \mathbf{S}_d$. To put it all together, $\sum_d \alpha_{nd} \mathbf{S}_d = \sum_d (\alpha_{nd} - m_d) \mathbf{S}_d + m_d \mathbf{S}_d, \forall n \leq N$ which clearly holds with $\beta_{nd} \geq 0, \forall n, d$. It is easy to see that there exists infinitely many choices for b_S , and thus β_n . One can simply set $b_S = \sum_d (m_d - \varepsilon_d) \mathbf{S}_d$ and $\beta_{nd} = \alpha_{nd} - (m_d - \varepsilon_d)$, where $\varepsilon_d \in \mathbf{R}_+, \forall d \in D$. Therefore, the final representation for the data sample Y_n is $\psi_n = \beta_n \mathbf{S} + b_S = \text{ReLU}(\beta'_n)\mathbf{S} + b_S$.

Obviously, we do not know α_n , thus β_n . However, just like training a network to output α_n , we can train a network to directly output β_n and set the bias term b_S as a learnable parameter to be learned from the data during the training process. The above derivation shows that, by using $\psi_n = \beta_n \mathbf{S} + b_S$ we do not lose any expressive power and still represent any sample as accurately as the common formula employed in literature, $\psi_n = \alpha_n \mathbf{S}$.

It may be distributed unchanged freely in print or electronic forms.



Figure 1: Our lifter architecture combines context vector with the estimated keypoints to produce improved pose estimates.

2 Lifter Network

The lifter network is a simple MLP with dedicated camera pose estimation as well as canonical shape head. CAN: END-TO-END MULTI-CATEGORY 3D POSE AND SHAPE



Figure 2: Two car samples from the validation set in (a) and (b). The images are the percategory closest samples in the latent space. The 3D reconstructions are from the latent code of the corresponding car mapped into different categories. Thus 3D shapes are <u>not</u> of the images but rather decoding of the latent codes of the cars. Elongation (height/width ratio) is translated across categories.

3 Additional Results

3.1 Disentanglement

We further demonstrate the properties of the latent space of the proposed formulation in Fig. 2. It can be seen that the latent space codes translate across object categories and encode geometric properties such as elongation. Moreover, the estimated 3D shapes of different categories from the same latent code match closely with the images that reside close to each other in the latent space. This proves our method's capability of successfully handling different classes.

We provide additional visual results on Pascal3D and Human3.6M datasets.



Figure 3: Visual results in Pascal3d dataset. The latent space of the proposed method can be decoded to produce 3D structure of any object category. The multi-category core of our method enables extraction of extra-categorical geometric relationships.

3.2 Visual results

Apart from the visuals provided in the paper, here we show additional results.



Figure 4: Visual results in Human3.6m dataset.



Figure 5: Visual results in Pascal3D dataset. It can be seen that our method produces very accurate results even in extremely difficult samples. For example, the car is seen from directly behind, yet our method recovers pose and shape very accurately.