Critical Research Analysis 2

Author: Jaskirat Singh
jaskirat.singh@anu.edu.au

1 Canonical Surface Mapping via Geometric Cycle Consistency. [Link]

Summary. The paper tackles the problem of mapping image pixels to a category-specific canonical 3D model. Unlike previous methods, the proposed approach requires minimum supervision as it utilizes the geometric consistency between the original pixels and the projection of the learned 3D surface back to the image plane. Kulkarni *et al.* [5] also incorporate a multi-pose prediction pipeline which solves the pose prediction and the canonical surface mapping problem in an end to end manner.

Strengths. 1. Learning without pose supervision: The proposed approach brilliantly alleviates the need for camera pose annotations by proposing to incorporate a multi-hypothesis pose prediction model in the overall geometric cycle pipeline. Furthermore, the authors use a diversity loss which prevents the problem of mode-collapse in pose prediction while encouraging exploration.

2. Intuitive Examples: The authors consistently provide intuitive examples in order to motivate the concepts introduced in the paper. *e.g.* why use visibility constraints? mask re-projection loss etc.

Weakness. 1. Visibility Constraints: **First,** The simultaneous application of L_{vis} and L_{cyc} can lead to opposing gradients for pixels $\mathbf{p}$ mapping to self-occluded regions on the 3D template. *e.g.* lets say for a front facing bird, $\mathbf{p}$ is on the beak, whereas the 3D location $\phi(C[\mathbf{p}])$ corresponds to the tail, with $\overline{\mathbf{p}} \approx \mathbf{p}$. Here, a gradient based L_{cyc} would like the 3D mapping to remain the same, despite the self-occluded nature of the projection. We hence suggest to use a masked cyclic loss as follows,

$$\text{Used in paper:} \quad L_{cyc} = \sum_{p \in I_f} ||\mathbf{p} - \overline{\mathbf{p}}||_2^2 \qquad \text{Suggested:} \quad L_{cyc} = \sum_{p \in I_f} \mathbb{1}_{[z_p - D_\pi[\overline{p}] < \epsilon]} \; ||\mathbf{p} - \overline{\mathbf{p}}||_2^2$$

Second, L_{vis} only discourages mapping to self-occluded 3D points. However, assuming a good enough NMR [3], a much stronger supervision for 3D mapping can be obtained by enforcing the z coordinate of the 3D point to be similar to the depth map at input pixel $D_\pi[p]$ (not $D_\pi[\overline{p}]$). We thus propose to use a depth consistency loss defined as: $L_{depth} = ||\; z_p - D_\pi[p] \;||_2$.

2. Confidence measure for APK: The confidence in keypoint transfer is measured as the inverse of the distance of between the source and target keypoint on the 3D template. However, for the same error in prediction, we would like the confidence to be high for regions on 3D model with relatively few keypoints (e.g. a birds tail), whereas the confidence should be relatively low for regions with high density of keypoints (e.g. a birds face with closely placed eyes and beak). We thus propose to model the confidence as the softmax over inverse distances computed from the K nearest source keypoints.

3. Synthetic data for better evaluation: The paper mentions unavailability of ground truth templates as a hindrance in directly evaluating model predictions. However, this could have been easily overcome by using synthetic datasets like [7], to evaluate the correctness of the learned 3D shape.

Opportunities. Adding to suggestions above, this work can be extended to instances with deformable shapes by learning a parameterizable 3D model. We also suggest introducing self-supervision through videos with an optical flow consistency loss between $\mathbf{p}$ from current frame and $\overline{\mathbf{p}}$ for the next frame.

Broader Impact. In addition to fostering development of graphical applications which use 3D understanding for photo-editing purposes, the proposed work has great potential to improve the safety of self-driving cars by estimating the 3D pose of the nearby vehicles. Since, the exact 3D model of the (nearby) vehicles is not required, a canonical surface mapping to a standard 3D car template, should provide sufficient information to estimate the 3D pose of a nearby vehicle.

2 PifPaf: Composite Fields for Human Pose Estimation. [Link]

Summary. The paper [4] tackles the problem of multi-person 2D human pose estimation for images with low resolution. The proposed method follows a bottom-up approach which uses Part Intensity Fields to predict the body part locations (with their relative scales), and a Part Association Field to learn the association between the predicted joints. Finally, Kreiss *et al.* [4] use the Laplace loss as an adaptive regression strategy to reduce the localization error for joints with lower scale (σ) / spread (b).

Strengths. 1. Novelty: The paper through Fig. 6 (top row, waving gesture) successfully demonstrates the practical importance of requiring accurate pose estimation while dealing with low resolution images. Furthermore, the proposed method, by learning scale parameters for both joint location and associations, is able to adopt a novel loss which reduces localization errors at smaller scales.

2. High quality visualizations: Qualitatively comparing pose prediction results, especially in overcrowded scenes, can be quite challenging. Nevertheless, the paper presents several high quality visualizations (refer Fig. 6,7,8) which crisply demonstrate the merits of the proposed method.

Weakness and Opportunities. 1. Clarity: **First,** the authors fail to explain the difference between the spread b and the scale σ learned through Part Intensity Fields. This is especially confusing, as both σ and b are used, as a measure of the scale of a joint, while computing the smooth-L1 and Laplace loss respectively. **Second,** the paper does not provide an adequate explanation for the significance of the width parameters (b_1, b_2) in the Part Association Fields and their difference with the corresponding spread parameters from Part Intensity Fields. Moreover, it seems that width parameters in PAF are not required and can be replaced by the spread parameter from the PIF of the first joint in Eq. 3.

2. Part Intensity Fields: The expression for confidence map $f(x, y)$ in Eq. 3 can be understood as a Gaussian Mixture Model with number of clusters equal to the number of pixels, cluster probabilities p_c^{ij}, cluster means (p_x^{ij}, p_y^{ij}) and standard deviation p_σ^{ij}. However, there are couple of problems with the current formulation. **First**, the expression for $f(x, y)$ in Eq. 3 should be normalized as $\sum_{ij} p_c^{ij}$ need not sum to one. Otherwise, it contradicts the heatmap range shown in Fig. 3c. **Second,** the authors currently assume a per-pixel spherical Gaussian distribution with $\Sigma = \textbf{diag}(p_\sigma^{ij}, p_\sigma^{ij})$. That is, they assume that the confidence scores vary uniformly along all directions from a given joint location. However, this is trivially incorrect. For e.g. we expect the confidence scores to reduce more slowly along the direction of the arm for a shoulder joint. A better confidence map can thus be learned by estimating three scale parameters (instead of one) to form a general representation for Σ^{ij}.

$$\text{Used in paper:} \quad \Sigma^{ij} = \begin{bmatrix} p_\sigma^{ij} & 0 \\ 0 & p_\sigma^{ij} \end{bmatrix} \qquad \text{Suggested:} \quad \Sigma^{ij} = \begin{bmatrix} p_{\sigma_x}^{ij} & p_{\sigma_{xy}}^{ij} \\ p_{\sigma_{xy}}^{ij} & p_{\sigma_y}^{ij} \end{bmatrix} \tag{1}$$

3. Inadequate Baselines: **First,** the paper uses only two non state of the art baselines to demonstrate the merits of the proposed approach. We thus, suggest the authors to include more recent works like [1, 2, 6, 8]. **Second,** the paper specifically mentions (Tab. 1) that Mask R-CNN was retrained for low resolution images. This raises some ambiguity on whether the same was done while reporting values for the OpenPose model, especially given the large skeletons detected by OpenPose in Fig. 6.

4. Abalation Studies: The paper fails to ascertain whether the proposed improvements result merely from adoption of a scale dependent regression loss versus the PifPaf formulation. Thus, the paper should include further experiments to shed light on the importance of predicting scale parameters through the part intensity and association fields. A possible ablation study should report results from MaskR-CNN with smooth-L1 loss where r_{smooth} is proportional to the area of the bounding box.

Opportunities for improvements. First, as mentioned above, we suggest to extend the proposed approach to model non-spherical Gaussian distributions while modeling the high-resolution confidence map. **Second,** the idea of scale-dependent loss functions could be extended to perform object detection and segmentation on low resolution images with Mask R-CNN, wherein the relative loss for bounding box corrections and class segmentation is scaled by the inverse of the bounding box area.

Broader Impact. As advertised in the paper, the proposed approach promises to have a huge impact in the development of self-driving cars. Accurate pose estimation at lower scales would enable the car to identify pedestrian gestures from a much greater distance, thus allowing more time for evasive actions if required. The ability to work with lower resolutions also implies faster processing times. Thus, this work would foster the development of real-time gesture recognition systems. A salient application would be violence detection in over-crowded areas through public surveillance videos.

References

[1] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5386–5395, 2020.

[2] T. Golda, T. Kalb, A. Schumann, and J. Beyerer. Human pose estimation for real-world crowded scenarios. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2019.

[3] H. Kato, Y. Ushiku, and T. Harada. Neural 3d mesh renderer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3907–3916, 2018.

[4] S. Kreiss, L. Bertoni, and A. Alahi. Pifpaf: Composite fields for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11977–11986, 2019.

[5] N. Kulkarni, A. Gupta, and S. Tulsiani. Canonical surface mapping via geometric cycle consistency. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2202–2211, 2019.

[6] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10863–10872, 2019.

[7] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *CVPR*, 2017.

[8] F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7093–7102, 2020.