Critical Research Analysis 1

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

1 Paper 1: SDC-Depth: Semantic Divide-and-Conquer Network for Monocular Depth Estimation. [Link]

Summary. The paper [10] tackles the problem of monocular depth estimation by incorporating scene priors from semantic and instance segmentation with a divide and conquer strategy. Wang *et al.* [10] decompose the original image into multiple semantic and instance segments, which have very consistent depth structures and thus, are easier inputs for depth estimation. Finally, they propose a depth aggregation pipeline which combines depths maps at category and instance levels with a bottom-up approach (instance $\rightarrow$ category $\rightarrow$ global) to output global depth predictions.

Strengths. 1. Novelity and leveraging segmentation datasets: **First**, the paper brilliantly utilizes the consistency of depth structures in low level semantic/instance segments to propose an end to end training pipeline for monocular depth estimation. **Second**, monocular depth estimation heavily relies on learning strong scene priors. Since, densely annotated depth datasets are limited, the proposed approach leverages large scale segmentation datasets [7] to improve semantic understanding.

2. Ablation Studies: The paper provides extensive ablation studies which highlight the importance of various model parts (Tab. 4) and the effect of segmentation data on depth prediction (Fig. 8).

Weakness. 1. Instance-wise depth transformation: The authors claim to learn the instance level transformations $\mathcal{H}^i(.)$ (from local to global depth), using only ROI Align features (that are local to the bounding box for each instance) and normalized bounding box coordinates. While this may suffice for certain cases, in general, accurate computation of this transformation would clearly require global context information. *e.g.* the depth offsets would be high in a large room and vice versa.

2. Relative weight-age between instance and category depths: The high value of the parameter ν in Eq. 1 is concerning because for $\nu = 10$ and reasonable p_i, mostly instance depth information would be retained after category-instance depth aggregation. For instance, the relative weight-age for the instance level depth map would be as high as $\approx 90\%$ $(5/6)$ for a very low $p_i = 0.5$. This means that, essentially, the proposed method uses instance depth maps for K object class categories and category-level depth maps for non-object classes. Thus, the paper should have an analysis measuring variation in performance metrics as ν is reduced, to shed more light on this problem.

Opportunities for improvements. Modelling uncertainty in instance depth prediction: As discussed above, the depth aggregation strategy for instance and category level depth maps has a very high weight-age ν for instance maps. Ideally, we would like to modify this weightage based on the model's confidence in predicted instance depths. The current model uses only p_i (probability of the i^{th} instance belonging to category c_i), as a measure of uncertainty in the instance-level depth maps. To this end, we suggest to model the parameters of the affine instance level transformation $\mathcal{G}^i(.)$ by a unimodal gaussian distribution, wherein the relative weightage of instance level depth map in Eq. 1, would be proportional to the prediction confidence (or inversely proportional to distribution entropy).

Broader Impact. While this work promises to be of key value in industrial computer vision applications like photo-editing and self-driving cars, its low-level societal impacts would extend far beyond that. For instance, this work can be used to improve the development of assistive technologies for the visually impaired *e.g.* obstacle detection [8] apps using depth estimation. Furthermore, this method can be easily incorporated into assistive scene-understanding devices, which already rely on segmentation results for semantic understanding.

2 Paper 2: 6D Camera Relocalization in Ambiguous Scenes via Continuous Multimodal Inference. [Link]

Summary. The paper [3] tackles the problem of measuring uncertainty for the 6 DoF camera pose estimation problem in ambiguous environments. The traditional regression strategy [2] doesn't provide a measure of confidence in the camera pose predictions and to this end, Bui *et al.* [3] propose to model the distribution for rotation and translation parameters using Bingham and Gaussian mixture models, respectively. Furthermore, [3] adopt the RWTA [9] loss function to avoid the problem of mode-collapse as seen in Mixture Density Networks (MDNs) [1] and MC-dropout [6] models.

Strengths. 1. Novelity: The paper successfully highlights an important problem on the limitation of previous camera-pose estimation methods in ambiguous scenes, containing multiple ground truth symmetries. In addition, the proposed training scheme allows for learning a diverse multi-modal distribution with deep neural networks, which can be applied to several other research applications.

2. Consistency of initial assumption with results: The authors, through Fig. 3, corroborate their initial assumption by showing high correlation between model uncertainty and prediction error. This consistency between results and assumptions, increases our confidence in the proposed method.

3. High quality visualizations: Despite its mathematically intense nature, the paper contains high quality figures (Fig. 1,4,6) that support the intuitions and concepts conveyed throughout the paper.

Weakness. 1. Inference: The method for pose prediction at test times (Section 5) is slightly flawed. The paper selects the optimal cluster as the mode of the cluster with the highest cluster probability $\pi_j(X_i, \Gamma)$. However, the probability of a cluster mode, in addition to the cluster probability $\pi_j(X_i, \Gamma)$, also heavily depends on the unimodal variance $|\Sigma_j|$ of the particular cluster. Thus, a possibly better metric for hypothesis selection, at inference time, could be given by,

$$\text{Used in paper:} \quad \arg\max_j \pi_j(X_i, \Gamma), \qquad \text{Suggested:} \quad \arg\max_j \frac{\pi_j(X_i, \Gamma)}{|\Sigma_j|^{1/2}}. \tag{1}$$

2. Evaluation metrics: The oracle error (Section 6) is biased to favor multi-modal distributions with high uncertainty, *e.g.* oracle accuracy would be high even if the RWTA method has high uncertainty for modes closest to the ground truth. Also, instead of accuracy with respect to the ground truth, it would be better to predict the correctness of a pose hypothesis, by computing the similarity between the query image and the image captured from the estimated pose, in the latent feature space (Fig. 2).

3. Concerns about Table 4: The values reported in Tab. 4 would vary highly with the 3D scene and number of ground truth symmetries. For instance, for the dining table dataset (Fig. 4) with only 2 ground truth symmetries, the proposed method would have a very low (≈ 0.04) fraction of correctly predicted modes. This would make it hard to demonstrate the differences between discussed methods. Instead, we suggest that if the number of ground truth symmteries is G, the % correctly predicted modes be computed on only the top G modes (ranked by uncertainity) predicted by the model.

4. RWTA for non-ambiguous scenes: The paper trains only unimodal and BMDN (practically unimodal) methods for the non-ambiguous dataset. Since, in practical applications, the ambiguity is not predetermined, it would be interesting to see if the RWTA method does at least as well.

Opportunities for improvements. First, As mentioned in Section 6.1 and Fig. 3, the model uncertainty shows high correlation with the prediction error. We can use this result to adopt a hard sample mining approach similar to Ding *et al.* [4] and increase the weightage of samples $\mathcal{X}_i$ with high entropy (refer Eq. 15) in the batch computation of RWTA loss. This would make the model focus more on hard examples during training. **Second**, the camera pose distribution shows high uncertainty when conditioned on a single query image. We thus propose to use sequential learning and recurrent neural networks to reduce this uncertainty based on trajectories leading to the ambiguous query.

Broader Impact. The proposed work promises significant impact in the thriving computer vision industry focusing on socially high demand applications like augmented reality, human computer guidance and robot guidance in an indoor environment with high ambiguity. For instance, in the social healthcare industry, a caretaker robot [5] must learn to localize its position in possibly ambiguous indoor environments. Recognition of samples having low certainty can help indicate to the robot to look for additional cues to determine its current location. Similarly, a self-driving car facing uncertainty in camera relocalization could rely on other sensory measurements like lidar, for determining its current location in the 3D scene.

References

[1] C. M. Bishop. Mixture density networks. 1994.

[2] E. Brachmann and C. Rother. Learning less is more-6d camera localization via 3d surface regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4654–4662, 2018.

[3] M. Bui, T. Birdal, H. Deng, S. Albarqouni, L. Guibas, S. Ilic, and N. Navab. 6d camera relocalization in ambiguous scenes via continuous multimodal inference. *arXiv preprint arXiv:2004.04807*, 2020.

[4] M. Ding, Z. Wang, J. Sun, J. Shi, and P. Luo. Camnet: Coarse-to-fine retrieval for camera re-localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2871–2880, 2019.

[5] J. Hoefinghoff, A. Rosenthal-von Der Pütten, J. Pauli, and N. Krämer. "yes dear, that belongs into the shelf!"-exploratory studies with elderly people who learn to train an adaptive robot companion. In *International Conference on Social Robotics*, pages 235–244. Springer, 2015.

[6] A. Kendall and R. Cipolla. Modelling uncertainty in deep learning for camera relocalization. In *2016 IEEE international conference on Robotics and Automation (ICRA)*, pages 4762–4769. IEEE, 2016.

[7] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[8] M. Mancini, G. Costante, P. Valigi, and T. A. Ciarfuglia. Fast robust monocular depth estimation for obstacle detection with fully convolutional networks. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4296–4303. IEEE, 2016.

[9] C. Rupprecht, I. Laina, R. DiPietro, M. Baust, F. Tombari, N. Navab, and G. D. Hager. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3591–3600, 2017.

[10] L. Wang, J. Zhang, O. Wang, Z. Lin, and H. Lu. Sdc-depth: Semantic divide-and-conquer network for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 541–550, 2020.