VEHICLE POSE ESTIMATION USING MASK MATCHING

Qingnan Li[1] *Ruimin Hu*[1,2,3*†] *Yu Chen*[1] *Yixin Chen*

[1]National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan 430072, China
[2]Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan 430072, China
[3]Collaborative Innovation Center of Geospatial Technology, Wuhan 430079, China

ABSTRACT

In this paper, we present a conceptually novel framework for vehicle pose estimation from given RGB images. Our approach extends Mask R-CNN by adding two branches for coarse viewpoint estimation and keypoint detection, in parallel with the existing branches for mask segmentation and 2D object detection in the training stage. Capitalizing on the estimated mask and the mask renderings from ShapeNet in the inference stage, we propose a mask optimization scheme to recover the vehicle poses from 2D-3D correspondences. Then, we enforce geometric constraint on these vehicle poses in a coarse-to-fine hybrid approach for robustness. Experimentally, our framework outperforms the state-of-the-art approaches on the very challenging PASCAL3D+ dataset.

***Index Terms*—** Multi-task R-CNN, Viewpoint, Keypoint, 3D Model, Mask

1. INTRODUCTION

Pose estimation has been widely studied over a long period of time because of its importance in a variety of applications such as robotic manipulation, automatic driving, etc. Recently, impressive work [1][2] in pose estimation are based on the CNN architecture. These methods estimate the pose using a PnP algorithm [3] that requires the correct detection of object 2D keypoints or 2D projections of the objects 3D bounding box corners. However, localizing them with a convnet sometimes leads to false detections since some of vehicle parts have considerably same characteristics, such as right wheels or left wheels, right headlight or left headlight, etc. Meanwhile, the pose optimization sometimes fails because of the huge shape variation, as shown in Figure 1.

For the symmetry of vehicle poses, we observe an interesting phenomenon. Given an image, a hidden keypoint corresponds to a convolutional response map with higher entropy of information than the visible keypoint. Based on this maximum entropy principle, we enforce geometric constraint on the pose estimation for robustness. Another issue is the variability of vehicle shapes. Since the 2D-3D correspondences exist, the numerous renderings under multiple poses from ShapeNet [4] provide an effective prior for 2D vehicle shapes. Based on this useful prior, we propose a multi-stage mask optimization scheme to recover the vehicle poses.

Fig. 1. Similar characteristics at different locations (left) and the shape variation (right). The above images are from PASCAL3D+ dataset.

In the next section, related work is reviewed. We present our proposed method in Section 3 and evaluate our performance on the very challenging PASCAL3D+ dataset [5] in Section 4.

2. RELATED WORK

R-CNN: The CNN based networks have proven their effectiveness in many computer vision tasks. Due to the success of CNN, R-CNN [6] is proposed, which attends a large amounts of candidate object regions and provides according confidence score of interest regions computed on deep feature maps. Faster R-CNN [7] is the extension of R-CNN that learns the attention mechanism with a Region Proposal Network (RPN). Mask R-CNN [8] extends Faster R-CNN by adding the mask branch for predicting an object mask. In this work, Mask R-CNN is extended by our Multi-task R-CNN that provides vehicle detections, binary masks, coarse viewpoints and keypoints.

*National Natural Science Foundation of China (61671336, 61671332, U1736206)
†Hubei Province Technological Innovation Major Project (2017AAA123)

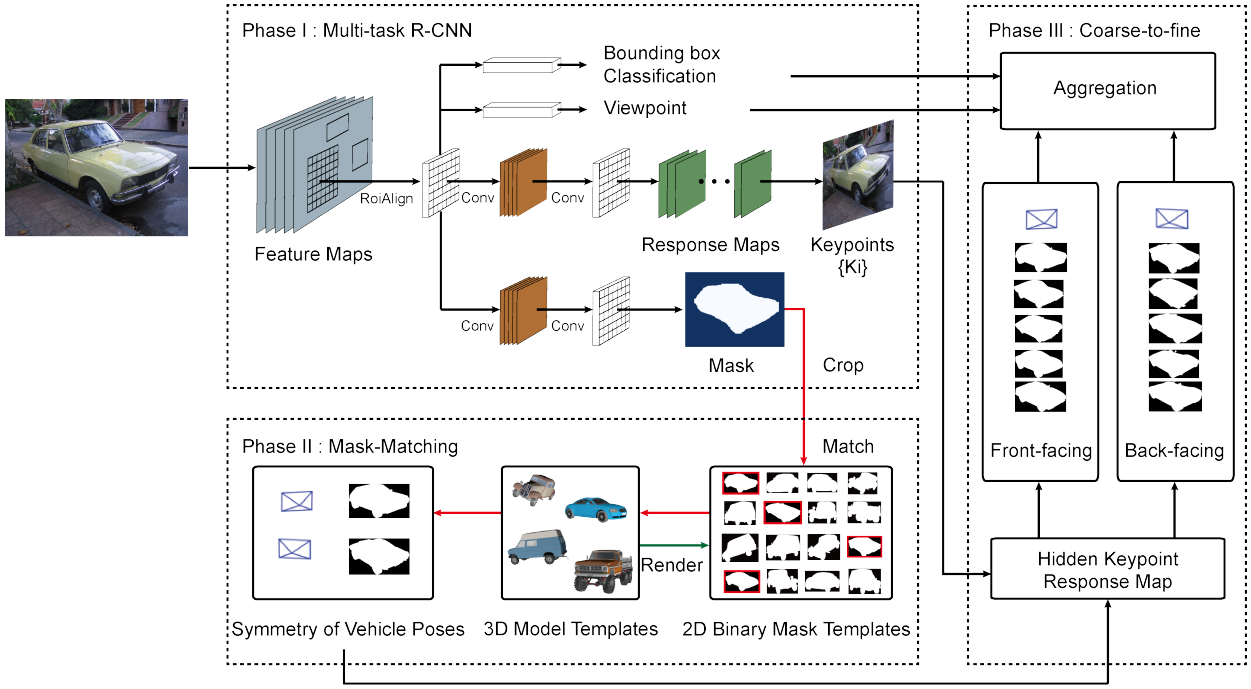

Fig. 2. Overview of our approach.

Pose Estimation: There are two ways to describe the pose of object: viewpoint based on global appearance and a fixed set of keypoints based on local appearance. Pose estimation is usually associated with viewpoint or keypoint based on R-CNN networks [1][2][9][10][11][12][13][14]. For viewpoint, these methods [9][12][13][14] generally divide the viewing sphere in several bins where each bin corresponds to a class. For keypoint, [10] introduces the viewpoint as the global appearance and infers the final keypoint locations based on 2D object analysis. To go further than 2D object reasoning, many methods [1][2][11] introduce 3D models and are able to give a detailed 3D object representation. These methods generally have a base R-CNN network providing object detections, semantic keypoints [1] or 2D projections of the objects 3D bounding box corners [2][11] for pose estimation and 3D model retrieval.

3. APPROACH

In this section, we describe the proposed approach for pose estimation from RGB images. Our approach is composed of three parts. First, the input image is passed through a Multi-task R-CNN network which outputs 2D detections, binary masks, coarse viewpoints and keypoints. The Multi-task R-CNN network architecture is detailed in Section 3.2. The second part is the inference using mask matching, detailed in Section 3.3. Additionally, associated with this module, the 3D model dataset and 2D binary mask dataset are detailed in Section 3.1. The last part is a coarse-to-fine procedure, mainly based on the maximum entropy principle, detailed in 3.4. The overview of our approach is illustrated in Figure 2.

3.1. Template Dataset

We use ShapeNet dataset [4] of M 3D models corresponding to various types of vehicles. For each 3D model m, its shape aligned in canonical view is denoted as $\bar{\mathbf{S}}_m^{3D}$. The set of 2D binary mask templates $\bar{\mathbf{T}}_m^{2D} = \{t_1, t_2, .., t_n\}$ associated to the 3D shape model $\bar{\mathbf{S}}_m^{3D}$ are renderings from diverse viewpoints by render pipeline [9], where t_n corresponds to the n^{th} binary mask template aligned in current viewpoint. By adopting a fine-grained (N=360) rendering viewpoints formulation, the 2D binary mask template dataset $\{\bar{\mathbf{T}}_m^{2D}\}_{m\in\{1,2,..,M\}}$ associated to the M 3D models totally has 2.7 million binary mask templates. Figure 2 shows some examples from 3D shape dataset $\{\bar{\mathbf{S}}_m^{3D}\}_{m\in\{1,2,..,M\}}$ and the 2D binary mask template dataset $\{\bar{\mathbf{T}}_m^{2D}\}_{m\in\{1,2,..,M\}}$.

3.2. Multi-task R-CNN Network

Our approach follows the spirit of Mask R-CNN, and makes two minor modifications to Mask R-CNN segmentation system. The first modification is the extension of Mask R-CNN by adding two branches for coarse viewpoint estimation and keypoint detection, in parallel with the existing branches for mask segmentation and 2D object detection. This is different from [10] in that global appearance is used in the inference stage, while our approach introduces viewpoint supervision

in the training stage for geometric constraint. Another modification is the shape constraint on keypoints detection. In contrast to the approach [1], the keypoints are limited within the mask instead of the bounding boxes.

Here, we train our Multi-task R-CNN on data from [5][15][16]. We detail all tasks of the Multi-task R-CNN network and the associated loss functions. Formally, during training, the network joint optimization minimizes the loss function $\mathcal{L} = \mathcal{L}_{det} + \mathcal{L}_{mask} + \mathcal{L}_{vp} + \mathcal{L}_{kp}$ on each sampled Region of Interest (RoI).

Detection loss: The detection loss $\mathcal{L}_{det}$ is associated with the classification $\mathcal{L}_{cls}$ and bounding box $\mathcal{L}_{box}$ that defined in Faster R-CNN [7].

Mask loss: The predicted mask representation from each RoI is parametrized as a single binary mask of resolution $m \times m$. We follow the definition of mask loss $\mathcal{L}_{mask}$ defined in Mask R-CNN [8].

Viewpoint loss: The predicted viewpoint representation is parametrized as a tuple of (θ, ϕ, ψ) of camera rotation parameters, where θ is azimuth angle, ϕ is elevation angle and ψ is cyclorotation angle. Compared with regression-based formulations [17], we adopt a fine-grained viewpoint classification formulation which typically has 360 discrete bins [9][10] for higher accuracy. We use a geometric structure aware loss function defined in [9] for geometric constraint.

Keypoint loss: The keypoint loss $\mathcal{L}_{kp}$ is related to the ground-truth class label p_c^*, defined as follows:

$$\mathcal{L}_{kp} = \lambda_{kp} \sum_{i=1}^{N_c} p_c^* P(K_i^* - K_i), \tag{1}$$

where P is the binary cross-entropy loss, and λ_{kp} is the regularization parameter of keypoint loss. The ground-truth class label p_c^* is 1 if the object proposal is vehicle and 0 otherwise. For each of N_c keypoints, the ground-truth keypoint label K_i^* is denoted as a one-hot $m \times m$ binary mask (m=56) and only a single pixel is denoted as foreground. Likewise, the predicted keypoint response map K_i is an m^2-way softmax output.

3.3. Mask Matching Inference

We use the Multi-task R-CNN network outputs, the 3D model dataset $\bar{\mathbf{S}}_m^{3D}$ and the mask template dataset $\bar{\mathbf{T}}_m^{2D}$ defined in 3.1 to recover the pose. Given a vehicle binary mask provided by Multi-task R-CNN network, the inference consists in two steps. In the first step, the mask output is cropped to its mask boundaries, denoted as $\bar{\mathbf{T}}_c$. In the second step, we minimizes the Euclidean distance E between the cropped mask $\bar{\mathbf{T}}_c$ and the 2D binary mask templates $\{\bar{\mathbf{T}}_m^{2D}\}_{m\in\{1,2,..,M\}}$:

$$t = \underset{m\in\{1,2,..,M\}}{\arg\min} \; E(\bar{\mathbf{T}}_m^{2D}, \bar{\mathbf{T}}_c). \tag{2}$$

Here, we choose the top k minimum distance as the coarse binary masks $\{t_k\}$. Although the resulting masks are considering similar, the symmetry of vehicle poses remains, as illustrated in Figure 3.

Fig. 3. The mask matching inference. The columns show the query RGB image from PASCAL3D+; the mask provided by our Multi-task R-CNN; the cropped mask $\bar{\mathbf{T}}_c$; the symmetry of vehicle poses in mask matching. The poses of the resulting matches are often far apart, due to the factor of vehicle symmetry (eg. front facing vs back facing car).

3.4. Coarse-to-fine Procedure

To address the symmetry of vehicle poses detailed in Section 3.3, a maximum entropy approach is proposed to separate the front-facing from the back-facing. We calculate the entropy of information for each of $\overline{N}_c$ keypoint response maps (left headlight, right headlight, left taillight, right taillight) and select the highest one as the response map of hidden keypoint:

$$k = \underset{\{K_i\}_{i\in\overline{N}_c}}{\arg\max} \; -\sum_{j=1}^{m^2} p_j^{K_i} \log p_j^{K_i}, \tag{3}$$

where k is the response map of hidden keypoint, and p_j is the value at each neuron contained in m^2-way softmax output K_i. Based on the maximum entropy principle, we enforce geometric constraint on the mask matches, and obtain a more fine-grained set of mask templates $\{t_k^*\}$.

Finally, we compare the coarse viewpoint proposal (provided by our Multi-task R-CNN directly) with the poses from refined mask templates $\{t_k^*\}$. If the confidence value of viewpoint proposal is less than 0.5, we choose a best pose from the above fine-grained mask templates $\{t_k^*\}$, which has the minimum distance to the cropped mask $\bar{\mathbf{T}}_c$ defined in Section 3.3. Otherwise, the viewpoint provided by Multi-task R-CNN is preferred.

4. EXPERIMENTS

Our experiments are divided into two parts. First, we evaluate our pose estimation approach on the challenging PASCAL3D+ dataset [5] (Section 4.1). Second, for the vehicle symmetry problem, we visualize each response map of the $\overline{N}_c$ keypoints, and analyze these maps (Section 4.2), which reflects the importance of maximum entropy in our multi-stage mask optimization scheme.

4.1. Viewpoint Estimation

We evaluate our viewpoint estimation approach without ground truth detections at runtime in different settings, start-

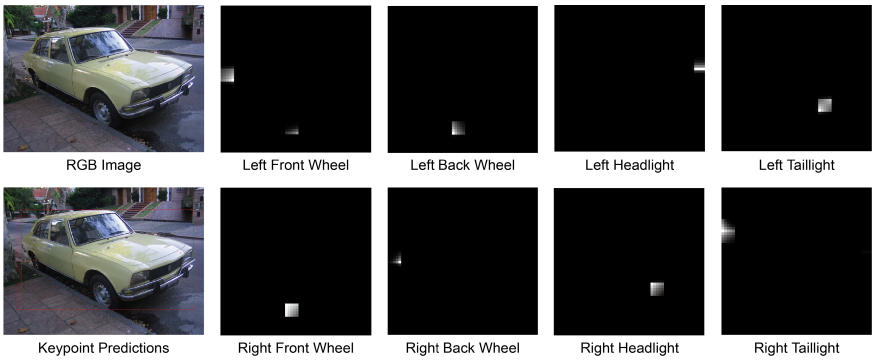

Fig. 4. Response Map Visualization. Each response map is an m^2-way output in our Multi-task R-CNN.

	$Acc_{\frac{\pi}{6}}$	$MedErr$
Su et al. [9]*	0.88	6.0
Tulsiani et al. [10]	0.89	9.1
Mousavian et al. [18]	0.90	5.8
Grabner et al. [2]	0.94	5.2
Pavlakos et al. [1]**	-	5.5
Ours - Multi-task R-CNN	0.88	7.2
Ours - Mask Optimization	**0.91**	**5.3**

Table 1. Viewpoint Estimation of Car. In the top part of the table, the approaches [2][9][10][18] use the ground truth detections on PASCAL3D+ dataset [5], while approaches in the bottom part estimate viewpoint without ground truth 2D detections. * The approach extends the training data by adding synthetic data. ** The ground truth 3D model must be known at the runtime.

ing from baseline Multi-task R-CNN version to Mask Optimization based version, following the evaluation protocol proposed in [10] to measure the rotation error

$$\Delta(R_{gt}, R_{pred}) = \frac{||log(R_{gt}^T R_{pred})||_F}{\sqrt{2}} \quad (4)$$

between the ground-truth R_{gt} and the estimated viewpoint rotation matrix R_{pred}. We report two metrics for evaluation: $Acc_{\frac{\pi}{6}}$ (the percentage of all viewpoint differences within $\frac{\pi}{6}$) and $MedErr$ (the median of all viewpoint differences that is robust to object symmetry). The AVP metric [5] is not applicable because 2D detections are not meaningful for our specific task. Quantitative results are presented in Table 1.

Without ground truth detections at runtime, our Multi-task R-CNN based approach outperforms [10] in $MedErr$ since the geometric constraint between viewpoints and keypoints are enforced in the training stage. Then, in order to address the symmetry of vehicle poses, our mask optimization scheme integrates the entropy of information for each of keypoint response maps. After a coarse-to-fine procedure, our approach outperforms the state-of-the-art [1] in $MedErr$ without the known 3D model.

4.2. Response Map Visualization and Analysis

The correct confirmation of hidden keypoint provides a solution to the symmetry of vehicle poses. We visualize the response map for each of $\overline{N}_c$ keypoints to reflect this nature. In order to get a better look at m^2-way response map K_i defined in Section 3.2, we multiply a factor α to the value at each neuron contained in K_i, where α equals $255/max(K_i)$.

Formally, the entropy of information reflects the disorder and randomness of neurons in the response map. In other words, the image brightness uniformity is a simple, elegant way to visualize the entropy of information, as shown in Figure 4. The second and third column in Figure 4 shows that our Multi-task R-CNN has its intrinsic difficulty for localizing those vehicle parts that have considerably same characteristics. As a comparison, we can observe that the hidden keypoint (left taillight) corresponds to a response map, in which the brightness is much more uniform than visible keypoints (left headlight, right headlight and right taillight). This confirms geometric constraint in our mask optimization scheme for robustness, and improves the performance on both $Acc_{\frac{\pi}{6}}$ and $MedErr$.

5. CONCLUSION

In this paper, we propose a novel approach for vehicle pose estimation from RGB images. Capitalizing on the 2D binary mask templates rendered from ShapeNet and the estimated mask provided by our Multi-task R-CNN, we propose a multi-stage mask optimization scheme to recover vehicle poses. To address the symmetry of vehicle poses, our approach integrates the entropy of information of each keypoint response map, and enforces geometric constraint on the vehicle poses. Experimentally, we demonstrate state-of-the-art results on Pascal3D+ dataset for pose estimation. Finally, we hope that our results motivate future research on pose estimation.

6. REFERENCES

[1] G. Pavlakos, X. Zhou, A. Chan, KG. Derpanis, and K. Daniilidis, "6-dof object pose from semantic keypoints," *IEEE International Conference on Robotics & Automation*, pp. 2011–2018, 2017.

[2] A. Grabner, PM. Roth, and V. Lepetit, "3d pose estimation and 3d model retrieval for objects in the wild," *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[3] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate o(n) solution to the pnp problem," *INTERNATIONAL JOURNAL OF COMPUTER VISION*, vol. 81, pp. 155–166, FEB. 2009.

[4] AX. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "Shapenet: An information-rich 3d model repository," *Technical Report arXiv:1512.03012 [cs.GR]*, Dec. 2015.

[5] Y. Xiang, R. Mottaghi, and S. Savarese, "Beyond pascal: A benchmark for 3d object detection in the wild," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2014.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, JUN. 2014.

[7] S. Ren, K. He, R. Girshick, and J Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 39, pp. 1137–1149, JUN. 2017.

[8] K. He, G. Gkioxari, P. Dollr, and R. Girshick, "Mask r-cnn," *IEEE transactions on pattern analysis and machine intelligence*, JUN. 2018.

[9] H. Su, CR. Qi, Y. Li, and LJ Guibas, "Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views," *IEEE International Conference on Computer Vision*, pp. 2686–2694, Dec. 2015.

[10] S. Tulsiani and J. Malik, "Viewpoints and keypoints," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1510–1519, Jun. 2015.

[11] J. Wu, T. Xue, JJ. Lim, Y. Tian, JB. Tenenbaum, A. Torralba, and WT. Freeman, "Single image 3d interpreter network," *European Conference on Computer Vision*, pp. 365–382, 2016.

[12] F. Massa, R. Marlet, and M. Aubry, "Crafting a multi-task cnn for viewpoint estimation," *British Machine Vision Conference*, 2016.

[13] P. Poirson, P. Ammirato, CY. Fu, W. Liu, J. Koeck, and AC. Berg, "Fast single shot detection and pose estimation," *International Conference on 3D Vision*, pp. 676–684, OCT. 2016.

[14] M. Elhoseiny, T. El-Gaaly, A. Bakry, and Ahmed Elgammal, "A comparative analysis and study of multiview cnn models for joint object categorization and pose estimation," *International Conference on Machine learning*, 2016.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, , and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, JUN. 2009.

[16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick, "Microsoft coco: Common objects in context," *13th European Conference on Computer Vision*, vol. 8693, pp. 740–755, SEP. 2014.

[17] F. Massa, M. Aubry, and R. Marlet, "Convolutional neural networks for joint object detection and pose estimation: A comparative study," *Computer Science*, pp. 412–417, 2015.

[18] A. Mousavian, D. Anguelov, J. Flynn, and J. Koeck, "3d bounding box estimation using deep learning and geometry," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5632–5640, JUL. 2017.