Supplementary Mateiral *for* 3D Key-points Estimation From Single-view RGB Images

Mohammad Zohaib[1,2][0000−0003−2259−4121], Matteo Taiana[1][0000−0003−1759−2447], Milind Gajanan Padalkar[1][0000−0002−0342−5448], and Alessio Del Bue[1][0000−0002−2262−4872]

[1] Pattern Analysis & Computer Vision (PAVIS), Italian Institute of Technology, Genoa, Italy
[2] Department of Marine, Electrical, Electronic and Telecommunications Engineering, University of Genoa, Italy
{mohammad.zohaib, matteo.taiana, milind.padalkar, alessio.delbue}@iit.it

1 Introduction

In this supplementary document, we present the qualitative results of the experiments presented in the main paper and evaluate the proposed network in different settings. The performance of the network is tested by removing the PWR module. It is found that there is a drop in the performance without this module. Moreover, the approach is evaluated for images with real backgrounds. The distribution of the angular distance error in pose estimation is computed, which verifies that the error remains within 0 to 5 degrees for most of the test samples.

2 Qualitative Results

In this section, we present qualitative results of the proposed approach for the ten other categories. The visualizations are depicted in Fig. 1. The first and the fourth row show the test images, the second and the fifth row present the estimated key-points on top of the original point clouds of the objects, and corresponding ground truth key-points are shown in the third and the sixth row. It can be observed that the key-points are estimated approximately on valid 3D positions and are in semantic order with respect to the ground truth key-points. Moreover, the proposed approach is able to predict 3D key-points for the occluded parts of the objects. For example, one leg of the table and the bed is not visible in the images because of self-occlusion. However, key-points are accurately estimated for them.

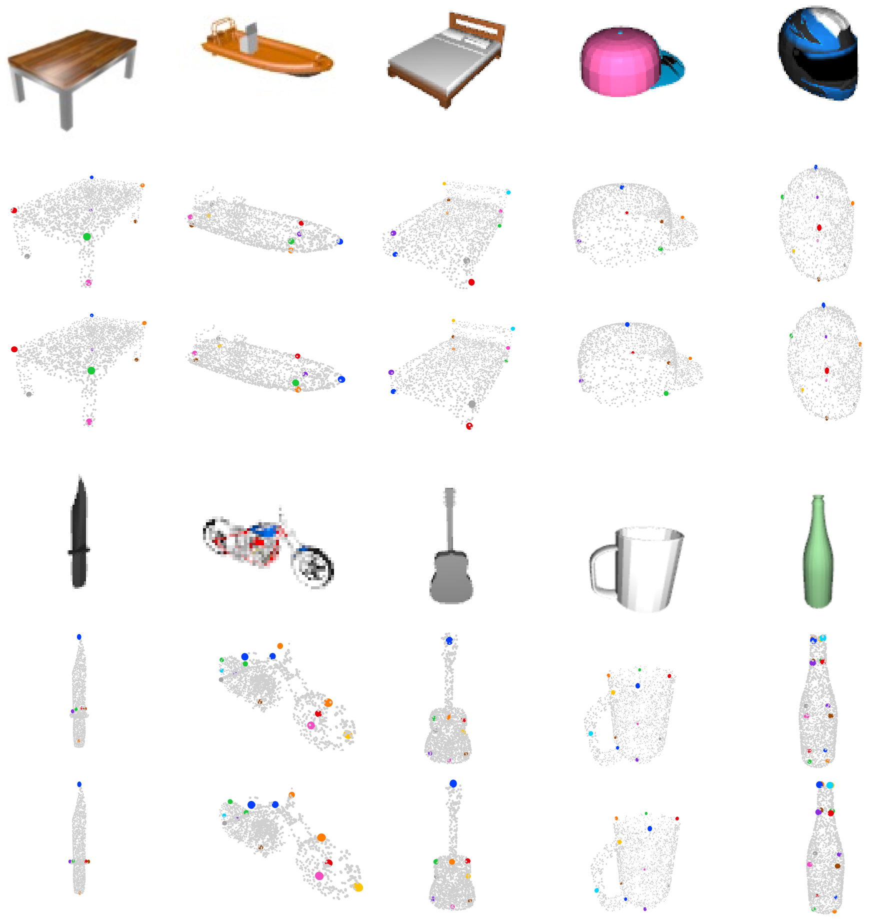


Fig. 1: Qualitative results of the proposed approach for other ten categories. Row (1, 4) show the input images, row (2, 5) and row (3, 6) present the corresponding estimated and ground truth key-points, respectively. It can be visualized that the proposed approach estimates a semantically ordered list of key-points even for the occluded parts of the objects.

3 Ablation study

3.1 Network without the PWR module

We revise the experiments for the transparent images of the three categories by removing the PWR module from the proposed network. The network can still attain better results than KP-Net [2]. However, the accuracy has reduced slightly in comparison with the complete network (with the PWR module). The comparison is shown in Tab. 1.

Table 1: Results for the architecture with and without the PWR module

Method	Cars		Planes		Chairs	
	Mean	Median	Mean	Median	Mean	Median
Ours with PWR	**5.190**	**2.073**	**3.257**	**2.053**	**10.732**	**4.096**
Ours without PWR	6.293	2.538	4.924	2.860	13.569	5.721

3.2 Test for realistic images

We evaluate our approach for images with real backgrounds that are taken from SUN dataset [1]. The angular distance error (in degrees) in pose estimation between two views of an object is depicted in Tab. 2.

Table 2: Results of our approach for images with a real background. The angular distance errors are calculated in degrees between the predicted and the ground truth rotation matrix using method 1 (Eq. 7 of main paper).

Method	Car		Airplane		Chair	
	Mean	Median	Mean	Median	Mean	Median
Ours with real background	41.47	12.84	51.01	29.27	70.782	61.52

Qualitative results of the proposed approach for realistic images are shown in Fig. 2. Column (a) shows the input images, whereas columns (b) and (c) depict the estimated and the corresponding ground truth key-points. The experiment shows that the results are not as impressive as they are in the case of synthetic images; RGB with a white background or RGBA with transparent background. It is due to the fact that the network could not separate the object from the background and hence estimates some key-points in the surrounding. This is our future task to improve the network for estimating more accurate 3D key-points from real background images.

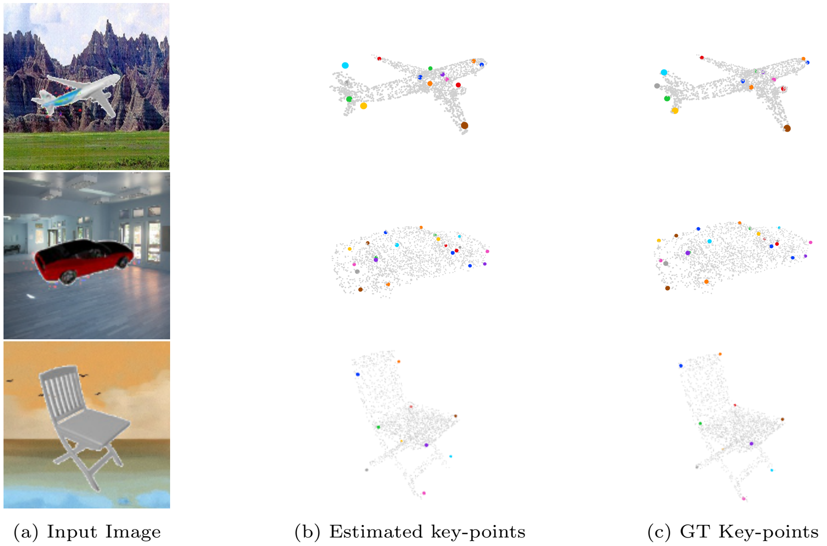


Fig. 2: Qualitative results of our approach for realistic images. (a) shows test images containing an object with a random background, (b) and (c) illustrate predicted and corresponding ground truth key-points on the object's point cloud, respectively.

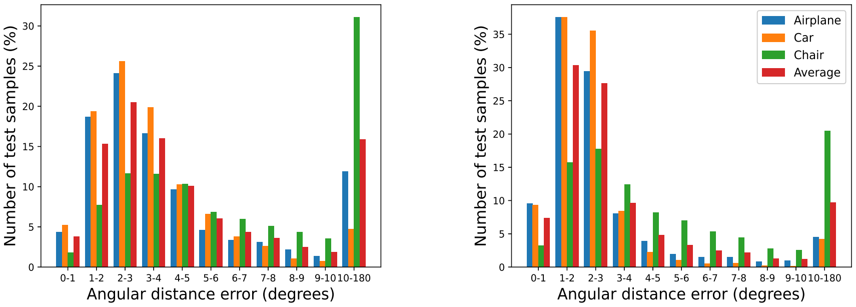
4 Additional Visualisations

4.1 Distribution of angular distance error

We present the distribution of the angular distance error computed between predicted and ground truth rotations (using method 1). We consider RGB (white background) and RGBA (transparent background) images. The corresponding histograms representing the computed distributions are shown in Fig. 3. It is observed that in both the cases RGB (Fig. 3a) and RGBA (Fig. 3b), the error for most of the test samples lies within 0 to 5 degrees. The error is less when RGBA images are used. Moreover, the error is comparatively high for the airplane category.

4.2 Evaluation of KP-Net for white background images

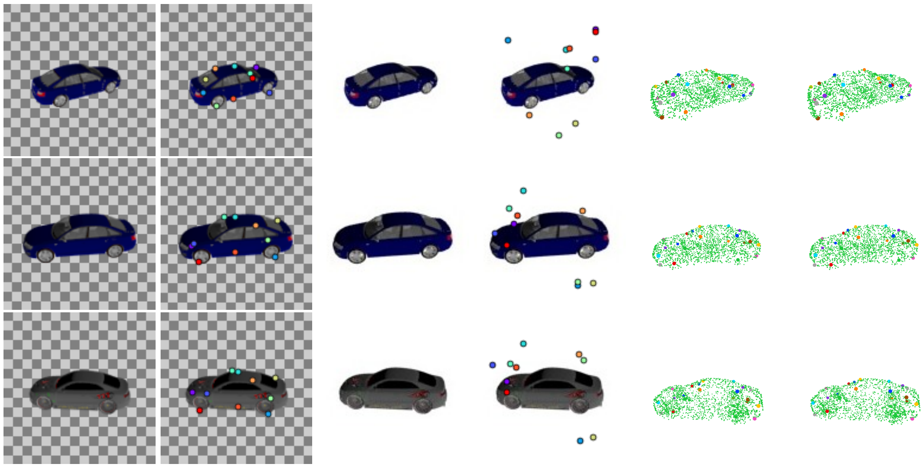
The KP-Net considers only RGBA images (transparent background). If we feed an RGB image (with white background), it first converts the image to RGBA by adding an additional channel and then processes it. We present results of the execution of KP-Net for different images in Fig. 4. We first execute KP-Net for RGBA image (Fig. 4a). The approach estimates key-points on the object (Fig. 4b). Then we execute it for RGB images (Fig. 4c). It fails to produce accurate key-points (Fig. 4d). The estimated key-points lie in the background. It



(a) RGB images

(b) RGBA images

Fig. 3: Distribution of angular distance error calculated between predicted and ground truth rotations computed using method 1. (a) and (b) show results for RGB (white background) and RGBA (transparent) images, respectively.



(a) RGBA (b) KP18 (c) RGB (d) KP18 (e) Our (f) Our GT

Fig. 4: The KP-Net produces poor results for RGB (white background) images. First, KP-Net is tested for RGBA (transparent) images (a), the corresponding predicted key-points are illustrated in (b). In second, it is tested for RGB (white background) images (c), the corresponding results of KP-Net are shown in (d). The estimations are incorrect. The same RGB images (c) are used to test the proposed approach, the results are depicted in (e). The ground truths are shown in (f).

is because the approach is trained for RGBA images, and it fails when tested for RGB images. In comparison, we execute our approach for the same RGB images

(Fig. 4c), it estimates accurate key-pints (Fig. 4e) which can be compared with the ground truth key-points (Fig. 4f).

References

1. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A. and Torralba, A..: Sun database: Large-scale scene recognition from abbey to zoo. CVPR. pp. 3485-3492 (2010)
2. Suwajanakorn, S., Snavely, N., Tompson, J., Norouzi, M.: Discovery of latent 3d keypoints via end-to-end geometric reasoning. NeurIPS. (2018)