Enhancing 6DoF Pose and Focal Length Estimation from Uncontrolled RGB Images for Robotics Vision

Mayura Manawadu[1] and Soon-Yong Park[1]

***Abstract*— Accurate 6DoF pose estimation is an important topic in robotics applications, from interactive systems to autonomous navigation and manipulation in augmented reality environments. Previous studies, that rely on single RGB image captured in uncontrolled environments often struggle to accurately estimate both the camera's internal focal length and the object's external pose parameters, primarily due to the inherent ambiguity of the perspective projection parameters of the pinhole camera model. Addressing this challenge, our study presents a two-stage approach by decoupling two projection related parameters by employing a render and compare strategy. Initially, we fix the z-axis translation (t_z) to an arbitrary value, effectively estimating the other pose parameters and focal length, and achieving accurate results when depth is assumed to be fixed. Subsequently, we predict all parameters in the second stage, enhancing the method's adaptability and accuracy by keeping the scale of the focal length to the object depth. This approach significantly overcomes projection scale ambiguity, devising improvements over existing methods. Both quantitative and qualitative results demonstrate the validity of presented approach showcasing its applicability for diverse robotics applications where accurate pose estimation is critical, yet camera metadata is either unreliable or unavailable.**

I. INTRODUCTION

6DoF pose estimation is a fundamental topic in computer vision which identifies object 3D position and orientation using translation and rotation matrices. It's importance is evident in many applications ranging from robotics [1], [2], Extended Reality (XR) [3], [4] and Simultaneous Localization and Mapping (SLAM) [5], [6]. Despite the existence of advanced techniques for 6DoF pose estimation from single RGB image [7], [3], a notable research gap persists in accurately estimating poses from images lacking metadata such as focal length. While cameras typically provide focal length metadata, there are numerous scenarios in robotics where this information may be absent, subject to change, or unreliable. These situations may include environments like disaster response zones, outer space missions, and underwater exploration, where conditions can significantly alter camera functionality or where pre-set metadata might not apply.

While Focalpose [8] presents a 6DoF pose estimation technique from in-the-wild obtained single RGB image without metadata like focal length, there exists an ambiguity in the predictions. This arises because of attempting a simultaneous prediction of the focal length and the translation along the Z-axis which influence the scaling of objects within the image. This method, employing a render-and-compare strategy, projects 3D pose estimations of objects onto 2D images to minimize projection error. However, the inherent coupling of internal focal length and external Z-axis translation in the perspective projection of the pinhole camera model complicates this process, as these parameters have opposite effects on projection scale. Despite a disentangled loss function aiming to separately adjust focal length and pose predictions, their update mechanisms remain coupled, leading to an inherent ambiguity which results in multiple solutions. Motivated by this work, a two stage approach is presented by decoupling the focal length and z-axis translation t_z predictions. In the first stage , we fix the z-axis translation to an arbitrary constant and predict all the other parameters. The outputs from the first stage, where the values are with respective to fixed t_z are used to initialize the values for the second stage. In the second stage, we predict all the 6DoF pose parameters except focal length which is scaled by predicted t_z. This two-stage methodology significantly mitigates the projection scale ambiguity inherent in single-image pose estimation tasks, and yeilds more accurate and reliable pose and focal length estimations.

In our evaluation, we observed a significant decrease of 18.04% in average median projection error, and 29.52% decrease in average median transformation error (Pose). These results demonstrate the method's effectiveness for robotics applications that require precise 6DoF pose and focal length estimations from single RGB image without metadata.

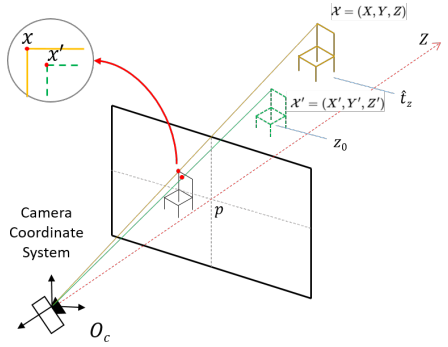

Fig. 1: Fixing t_z to an arbitrary constant to prevent the ambiguity between simultaneous prediction of focal length and t_z.

[1]Mayura Manawadu and Soon-Yong Park are with the School of Electronic and Electrical Engineering, Kyungpook National University, Daegu, South Korea mayuramanawadu@knu.ac.kr, sypark@knu.ac.kr

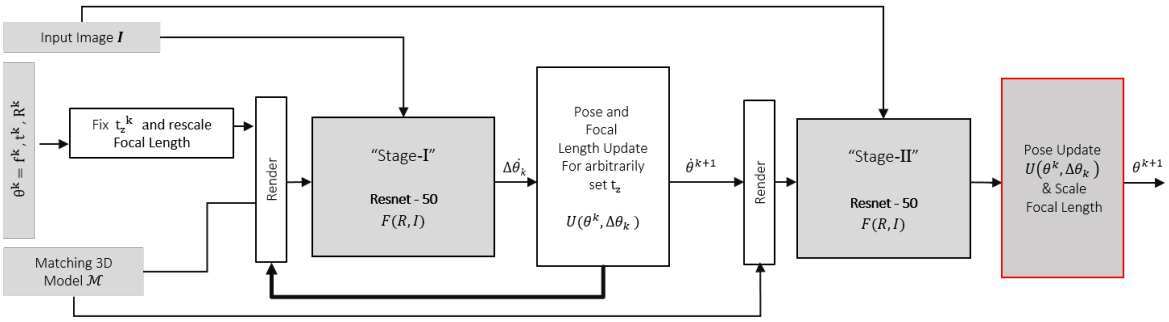

Fig. 2: Pose and Focal Length Estimator Network using a two stage approach to predict 6DoF and Focal Length.

II. METHODOLOGY

A. Motivation

Our methodology was motivated by the 2D image formation based on the perspective projection model (PPM) of a pin-hole camera setting. This model is comprised of two key parameters which affect the scale of an object's projection: the focal length and the object's translation along the Z coordinate in the camera's coordinate system. This setup introduces ambiguity in predicting the pose when these parameters are estimated simultaneously.

Our objective is to estimate the 6DoF pose from a single RGB image without metadata. As illustrated in Fig. 1, we need to determine the pose of the object such that its projection can be accurately rendered onto the RGB image when it is properly positioned within the camera's coordinate system. Conversely, given an object in the world coordinate system and a single RGB image, we seek to adjust the camera's position to ensure the object's accurate projection within the image.

Referring to Fig. 1, to project the model of the chair on to the image plane, it has to be positioned at $\mathbf{X_c} = (X_c, Y_c, Z_c)$ in the camera space, and it's resultant image coordinate will be x (for simplicity we will consider only x). The world coordinate system($\mathbf{X_w}$ - CAD Coordinates) is located at the origin and let's denote it by $\mathbf{X}_w$. If we consider the rotation components of extrinsic matrix to be $\mathbf{R} = [\mathbf{r_1}, \mathbf{r_2}, \mathbf{r_3}]^T$, the translation vector to be $\mathbf{t} = [t_x, t_y, t_z]^T$, and focal length by f, we can represent the relationship between x and X_c by the Eq. (1)

$$x = \frac{f}{\mathbf{r_3^T X}_w + t_z} X_c \tag{1}$$

Predicting the object's pose involves estimating the rotation matrix $\mathbf{R}$, translation vector t, and focal length f. In the equation above, t_z and f directly impact the projection scale, leading to ambiguity in their simultaneous prediction. To tackle this, we initially set t_z to a constant in the first stage and re-calibrate f accordingly.

Letting z_0 represent the fixed t_z, the change in image space from x to x' and CAD coordinates to $\mathbf{X'_c} = (X'_c, Y'_c, Z'_c)$ is illustrated in Eq. (2).

$$x' = \frac{f}{\mathbf{r_3^T X}_w + z_0} X'_c \tag{2}$$

To align x' with x after setting z_0, we re-calibrate f by the ratio of $x : x'$ (denoted S_f), as shown in Eq. (3):

$$S_f = x/x' = \frac{\mathbf{r_3^T X}_w + z_0}{\mathbf{r_3^T X}_w + t_z} \tag{3}$$

This re-calibration to align $x' \rightarrow x$ is given in Eq. (4). Here f' is the re-calibrated focal length ($f' = S_f f$). As there is no change done on t_x and t_y, $X_c = X'_c$ and $Y_c = Y'_c$.

$$x' \frac{x}{x'} = x' S_f = \frac{f S_f}{\mathbf{r_3^T X}_w + t_z} X_c = \frac{f'}{\mathbf{r_3^T X}_w + t_z} X_c \tag{4}$$

Our CAD models are computed as point clouds with the centroids at the origin. If we are to re-calibrate the focal lengths and do the computations for every point, it will be computationally expensive. Hence, by weak perspective projection [9] we set $r_3^T \mathbf{X}_w = 0$. Thus, after fixing t_z, the focal length adjustment simplifies to Eq. (5):

$$f' = f \frac{z_0}{t_z} \tag{5}$$

B. Approach Overview

Fig. 2 demonstrates a two-stage approach for predicting the 6DoF pose from an uncontrolled RGB image. In the first stage, t_z is fixed to an arbitrary constant, and other pose parameters and focal length are estimated. To offset the arbitrary setting, re-calibration of the focal length is necessary, as indicated by Eq. (5). Initially, a coarse estimation is performed, followed by iterative refinements. Outputs from the first stage are then utilized to initialize the values for the second stage, where t_z is predicted. In the second stage, instead of predicting the focal length anew, it is scaled based on the t_z update.

C. Stage-I : Prediction relative to fixed t_z

Inputs are an image and a 3D model with initial pose $\theta^k = (R^k, t_x^k, t_y^k, t_z^k, f^k)$, with t_z fixed at z_0 and focal length re-calibrated according to Eq. (5). After rendering the model with fixed t_z and adjusted focal length, a Resnet-50 network predicts the pose updates $\Delta\theta_k$. Eventhough a disentangled loss function has been proposed in [8] to decompose the pose and focal length updates, in their update rule for translation, focal length is taken in to consideration, which results in an ambiguity. Addressing this ambiguity in update, we fix

t_z in Stage-I, modifying the update rules $U = (\theta, \Delta\theta)$ for t_x and t_y as per Eq. (6) and Eq. (7), with updates ($\Delta\theta$) from network-1 outputs. Here v_x^k and v_y^k are the outputs from network-1 which are the values required for the update of t_x and t_y respectively during k iteration ($k > 1$ in refiner). In this stage, the update rule for rotation and focal length remains the same as [8].

$$t_x{}^{k+1} = \frac{v_x^k}{f^{k+1}} z_0 + t_x{}^k \tag{6}$$

$$t_y{}^{k+1} = \frac{v_y^k}{f^{k+1}} z_0 + t_y{}^k \tag{7}$$

The loss function $\mathcal{L}_{stage1}$, tailored for fixed t_z and re-calibrated focal length, is given by Eq. (8) to (10). Even though we have fixed a component of translation, we are estimating focal length at this stage. Hence for jointly learning the pose parameters and focal length, $\mathcal{L}_{pose}$ and $\mathcal{L}_{focal}$ are used. In Eq. (9), function D, computes the L_1 norm of the transformed points with given rotation and translation, between predicted and ground-truth values. $t_{xy} = (t_x, t_y, z_0)$ represents the translation with fixed t_z.

$$\mathcal{L}_{stage1}(\theta, \hat{\theta'}) = \alpha L_{\text{focal}} + L_{\text{pose}} \tag{8}$$

$$\begin{aligned} \mathcal{L}_{pose} = D(U(\theta^k, \{v_x^k, v_y^k, z_0, \hat{R}^k, \hat{v}_f^k\}), \hat{R}, \hat{t}) \\ + D(U(\theta^k, \{\hat{v}_x^k, \hat{v}_y^k, z_0, R^k, \hat{v}_f^k\}), \hat{R}, \hat{t}) \end{aligned} \tag{9}$$

Eq. (10) defines $\mathcal{L}_{focal}$ as the Huber Regression and disentangled re-projection loss, calculating the $L1$ norm of projected points. $\hat{f}'$ and $\hat{t}_{xy}$ denote the re-calibrated ground truth focal length and x, y translations respectively, with f being the predicted focal length.

$$\begin{aligned} \mathcal{L}_{focal} = \beta \mathcal{L}_H \left(f, \hat{f}'\right) + \\ \frac{1}{2}\mathcal{L}_{proj.}\left((R, t_{xy}, \hat{f}'), (\hat{R}, \hat{t}_{xy}, \hat{f}')\right) + \\ \frac{1}{2}\mathcal{L}_{proj.}\left((\hat{R}, \hat{t}_{xy}, f), (\hat{R}, \hat{t}_{xy}, \hat{f}')\right) \end{aligned} \tag{10}$$

D. Stage-II : Estimating 6DoF pose while scaling f

Utilizing Stage-I outputs relative to the fixed t_z, Stage-II directly predicts t_z using another Resnet-50, while scaling focal length according to the updated t_z. This stage omits refinement for direct pose prediction. Using just the a single pass in the network in second stage led to better results than iterative refinement, likely because the refiner can cause overfitting and make initial errors worse. When considering the update rules of the second stage, similar to previous, the rotation update remains the same. Update rule for t_z is given by Eq. (11). $t_z{}^{init}$ and f^{init} are the outputs from Stage-I.

$$t_z{}^{output} = v_z t_z{}^{init} \tag{11}$$

Here instead of predicting, we scale the focal length similar to relative $t_z{}^{output}$ update as in Eq. (12), thus eliminating the ambiguity.

$$f^{output} = f^{init}\left(\frac{t_z{}^{output}}{z_0}\right) \tag{12}$$

Then we update the x and y components of the translation as given by Eq. (13) and Eq. (14)

$$t_x{}^{output} = \left(\frac{v_x}{f^{output}} + \frac{t_x{}^{old}}{t_z{}^{init}}\right) t_z{}^{output} \tag{13}$$

$$t_y{}^{output} = \left(\frac{v_y}{f^{output}} + \frac{t_y{}^{old}}{t_z{}^{init}}\right) t_z{}^{output} \tag{14}$$

$$\begin{aligned} \mathcal{L}_{pose} = D(U(\theta^{old}, \{v_x, v_y, v_z, \hat{R}^k\}), \hat{R}, \hat{t}) \\ + D(U(\theta^{old}, \{\hat{v}_x^k, \hat{v}_y^k, v_z, \hat{R}^k\}), \hat{R}, \hat{t}) \\ + D(U(\theta^{old}, \{\hat{v}_x^k, \hat{v}_y^k, \hat{v_z}, R^k\}), \hat{R}, \hat{t}) \end{aligned} \tag{15}$$

The pose loss function L_{pose} given by Eq. (15), specified for this stage, concentrates solely on pose parameters excluding focal length estimation.

III. EXPERIMENTAL RESULTS

We evaluate our two-stage approach by comparing with FocalPose using the Pix3D dataset [10], showing marked improvements. We only used real dataset of Pix3D without synthetic data, due to the hardware constraints on training. Results in Table I shows our Stage-I (Ours-Fixed t_z) and Stage-II (Ours) performances versus FocalPose.

A. Quantitative Results

Our experiments, using metrics from [8], evaluate rotation, translation, pose, and projection via median reprojection errors ($Med.Err.$) and the percentage of images with reprojection errors below 0.1 ($Acc_{P_{0.1}}$) and 0.05 ($Acc_{P_{0.05}}$) of the image size. Pose error represents the error between 3D points of CAD model transformed using estimated rotation and translation with the ground truth in camera coordiante system. These metrics confirm our method's precision in aligning 3D models with 2D images.

Stage-I significantly outperforms FocalPose, demonstrating effectiveness with a fixed t_z. This stage is applicable where depth data is measurable. Stage-II, while not exceeding Stage-I's performance, improves upon FocalPose, supporting our method's capability for full 6DoF pose estimation including focal length.

Training was performed on an NVIDIA RTX 3090 GPU. All datasets were trained for 500 epochs, except for the chair dataset, which was trained for 200 epochs after noting performance stabilization at 150 epochs, ensuring efficient resource use while maintaining quality. Despite real data's noise, our method outperformed FocalPose in many metrics but faced challenges in projection accuracy for the Sofa dataset. This issue likely comes from real data's noise. However, better performance even on noisy real data indicates potential for enhanced accuracy on synthetic data with precise annotations.

TABLE I: Quantitative Comparison of Our Approach with FocalPose.

Dataset		Rotation				Translation	Pose	Focal	Projection		
		Med. Err.	Acc 30°	Acc 15°	Acc 5°	(Med. Err.)	(Med. Err.)	(Med. Err.)	Med. Err. 2	$\text{Acc}_{P_{0.1}}$	$\text{Acc}_{P_{0.05}}$
Pix3d Bed	FocalPose	0.436	53.68 %	32.11 %	3.16 %	0.251	0.202	0.222	0.132	41.05 %	13.16 %
	Ours-Fixed t_z	**0.389**	**62.11 %**	**37.89 %**	**6.32 %**	**0.019**	**0.044**	**0.064**	**0.104**	**47.37 %**	**20.53 %**
	Ours	**0.382**	**60.00 %**	**36.32 %**	**7.89 %**	**0.200**	**0.179**	**0.208**	**0.119**	**45.26 %**	**18.42 %**
Pix3d Sofa	FocalPose	0.236	79.78 %	56.77 %	10.39 %	0.230	0.153	0.208	0.057	74.77 %	43.04 %
	Ours-Fixed t_z	**0.134**	**94.07 %**	**80.37 %**	**30.56 %**	**0.012**	**0.017**	**0.038**	**0.038**	**87.04 %**	**65.37 %**
	Ours	**0.169**	**92.02 %**	**74.21 %**	**20.04 %**	**0.200**	**0.132**	**0.194**	**0.056**	**81.45 %**	41.19 %
Pix3d Table	FocalPose	0.762	36.75 %	17.38 %	1.71 %	0.503	0.312	0.323	0.204	19.09 %	3.70 %
	Ours-Fixed t_z	**0.500**	**51.28 %**	**27.07 %**	**3.70 %**	**0.021**	**0.053**	**0.075**	**0.136**	**38.46 %**	**15.38 %**
	Ours	**0.587**	**47.29 %**	**26.50 %**	**4.56 %**	**0.279**	**0.213**	**0.315**	**0.180**	**27.07 %**	**7.41 %**
Pix3d Chair	FocalPose	0.964	24.08 %	7.47 %	0.44 %	0.553	0.376	0.210	0.182	16.17 %	1.45 %
	Ours-Fixed t_z	**0.278**	**66.69 %**	**47.95 %**	**7.86 %**	**0.020**	**0.026**	**0.061**	**0.068**	**62.44 %**	**35.26 %**
	Ours	**0.288**	**66.35 %**	**44.96 %**	**7.40 %**	**0.216**	**0.146**	0.210	**0.096**	**51.56 %**	**20.96 %**

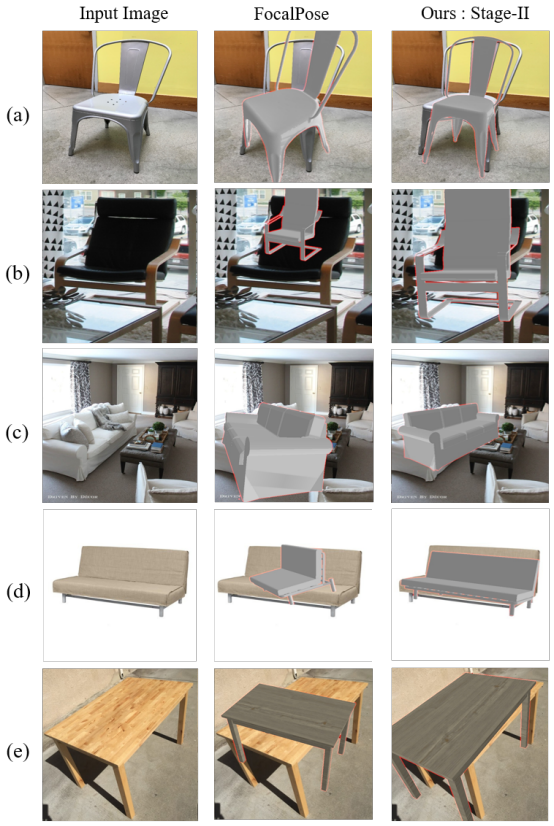

Fig. 3: Qualitative Comparison of Our Approach vs Focalpose.

Our approach shows notable improvements in key metrics as given in Table I. These are crucial for robotics, where accurate environmental understanding is necessary for tasks such as object manipulation and navigation. Better depth and focal length estimations enhance object manipulation and spatial awareness, while improvements in rotation and translation aid in precise pose estimation. These advancements enable more effective navigation and interaction within various environments, highlighting our approach's potential to boost robotics in settings where camera metadata is uncertain.

B. Qualitative Results

Fig. 3 visually illustrates our approach's performance, comparing with FocalPose. It shows instances where projection scales of rendered CAD models of our approach closely align with RGB images, reflecting accurate focal length and depth estimations. It can be observed that, improper projection scales exists in [8] which is due to ambiguity in simultaneous estimation of parameters. However, our results have room for improvements in rotation and translation, highlighting the potential advantages of integrating more synthetic data into our training process.

IV. CONCLUSION

In conclusion, this paper presents a novel two-stage approach to 6DoF pose estimation from single RGB image, effectively addressing the inherent scale ambiguity in simultaneous focal length and translation (t_z) estimation. By separating the estimation processes for t_z and focal length, we achieve enhanced performance, demonstrating significant improvements across a range of metrics. Our experiments on the real-world Pix3D dataset show the robustness and adaptability of our method. The qualitative evaluation further shows our approach's advantage in more accurately estimating pose and focal length. This research makes our method suitable for advanced robotics applications, promising advancements in autonomous navigation, object manipulation, and interaction within unstructured environments where camera metadata is is unreliable or unavailable.

REFERENCES

[1] S. Wang, J. Liu, Q. Lu, Z. Liu, Y. Zeng, D. Zhang, and B. Chen, "6d pose estimation for vision-guided robot grasping based on monocular camera," in *2023 6th International Conference on Robotics, Control and Automation Engineering (RCAE)*, pp. 13–17, 2023.

[2] J. Zhang, B. Yin, X. Xiao, and H. Yang, "3d detection and 6d pose estimation of texture-less objects for robot grasping," in *2021 6th International Conference on Control and Robotics Engineering (ICCRE)*, pp. 33–38, 2021.

[3] Y. Lu, S. Kourian, C. Salvaggio, C. Xu, and G. Lu, "Single image 3d vehicle pose estimation for augmented reality," in *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 1–5, IEEE, 2019.

[4] R. Hachiuma and H. Saito, "Recognition and pose estimation of primitive shapes from depth images for spatial augmented reality," in *2016 IEEE 2nd Workshop on Everyday Virtual Reality (WEVR)*, pp. 32–35, IEEE, 2016.

[5] X. Ruan, F. Wang, and J. Huang, "Relative pose estimation of visual slam based on convolutional neural networks," in *2019 Chinese Control Conference (CCC)*, pp. 8827–8832, IEEE, 2019.

[6] Z. Xiao, X. Wang, J. Wang, and Z. Wu, "Monocular orb slam based on initialization by marker pose estimation," in *2017 IEEE International Conference on Information and Automation (ICIA)*, pp. 678–682, IEEE, 2017.

[7] S.-Y. Park, C.-M. Son, W.-J. Jeong, and S. Park, "Relative pose estimation between image object and shapenet cad model for automatic 4-dof annotation," *Applied Sciences*, vol. 13, no. 2, p. 693, 2023.

[8] G. Ponimatkin, Y. Labbé, B. Russell, M. Aubry, and J. Sivic, "Focal length and object pose estimation via render and compare," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3825–3834, 2022.

[9] I. Shimshoni, R. Basri, and E. Rivlin, "A geometric interpretation of weak-perspective motion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 3, pp. 252–257, 1999.

[10] X. Sun, J. Wu, X. Zhang, Z. Zhang, C. Zhang, T. Xue, J. B. Tenenbaum, and W. T. Freeman, "Pix3d: Dataset and methods for single-image 3d shape modeling," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2974–2983, 2018.