AI637 spring 2024 final project report: Monocular 6D spacecraft pose estimation

Venkat Kalyanakumar Zhanpei Fang
Oregon State University, Corvallis, OR
{kalyanav, fangzha}@oregonstate.edu

Abstract

In this project we treat the problem of uncooperative 6D spacecraft pose estimation using monocular camera images. Building off of the unsupervised domain adaptation (UDA) approach of Pérez-Villar et al. [9], we try improving the keypoint localization using (1) replacing the 2D-2D mean-squared error (MSE) loss used to supervise the UDA model with an Adaptive Wing (AWing) loss which helps focus training on foreground over background pixels, and (2) replacing the two stacked hourglasses of the UDA model with a Cascaded Pyramid Network (CPN). We show that these modifications lead to improved performance on finding occluded and otherwise "hard" keypoints and therefore final pose score.

1. Introduction and background

The spacecraft pose estimation (SPE) problem is to compute the 6D pose of a non-cooperative target spacecraft with monocular cameras, assumed to be mounted on a small chaser spacecraft; from a single monocular image, get the transformation $T = [R|t] \in \mathbb{R}^{3\times4}$ where R is the rotation matrix and t the translation vector, relative to a canonical pose (Fig 1). This is important for space orbital missions which do on-orbit servicing and active debris removal, spaceflight capabilities which will become increasingly crucial in the next decade.

Challenges that are unique to this context, distinct from other 6D pose estimation problems, include the lack of real mission data and the hardware constraints of on-board deployment. Harsh lighting conditions in space lead to frequently occluded keypoints, caused by strong shadows and large variations in pose. Feature-based approaches do poorly in these variable illumination conditions, low SNR, and high contrast which is typically present in space imagery. Historically, the problem has been treated with two classes of models (Fig. 2): (1) direct end-to-end deep learning approaches and (2) hybrid modular approaches. The latter class of methods, which exploit the geometry of the problem to optimize pose from a correspondence set, tend to perform better than end-to-end approaches [1, 7], and so we focus on these in our study.

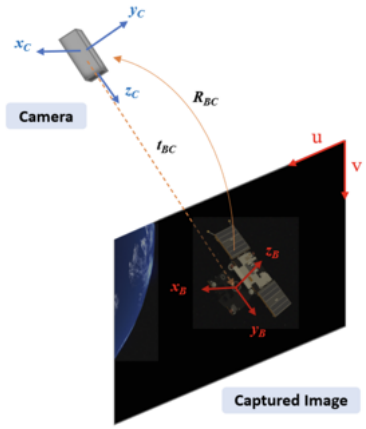

Figure 1. Spacecraft pose estimation is the problem of finding the relative position (t_{BC}) and orientation (R_{BC}) of a target spacecraft reference frame (B) shown in red, with respect to the camera reference frame (C) in blue, mounted on a chaser spacecraft. Figure from [7].

Hybrid modular approaches typically consist of three stages (Fig. 3): (1) a spacecraft detection/localization stage to crop the image, (2) a keypoint prediction stage which predicts 2D keypoint location of predefined 3D keypoints, and (3) pose computation to get pose from the 2D-3D correspondences. The first two stages are done using deep learning and the third typically uses a classical algorithm which does the outlier removal (e.g. RANSAC) necessary for the PnP (e.g. EPnP) solver and final pose refinement with optimization techniques.

The hybrid approach requires recovering 3D locations of keypoints with 3D reconstruction methods, which are used to construct secondary annotations (such as bounding boxes, keypoints, segmentation masks, ellipse heatmap annotations), so the lack of diverse annotated datasets is a major constraint, and the annotated datasets which do exist typically only contain one target spacecraft type. Many prior approaches try to exploit well-explored human pose methods for the spacecraft problem, which is a domain gap to

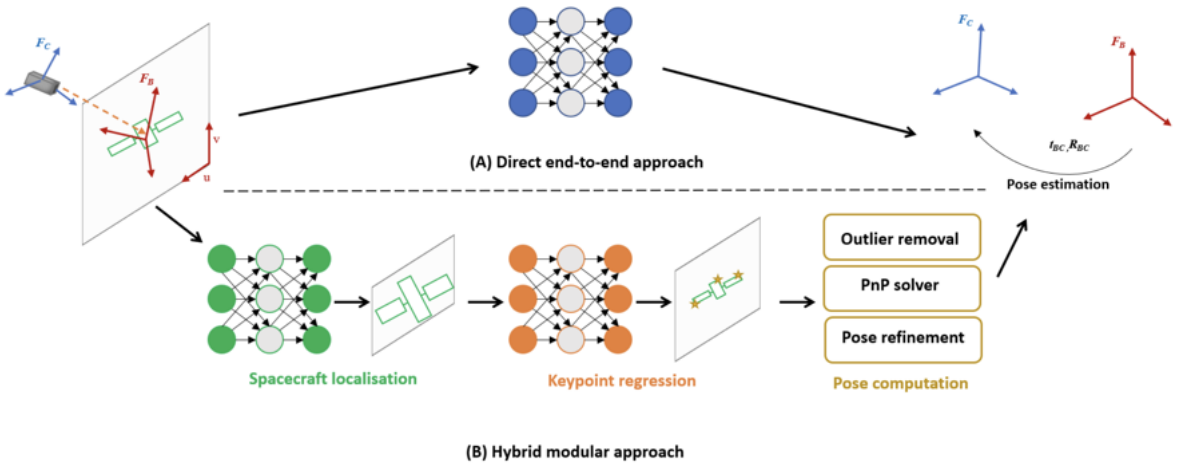

Figure 2. The two general classes of solutions to the spacecraft pose estimation problem. From [7].

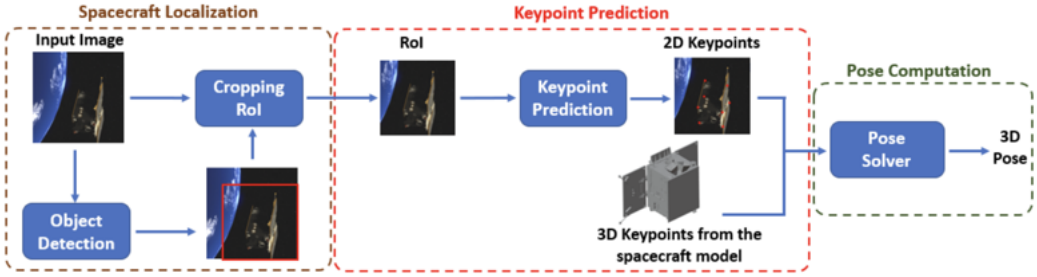

Figure 3. Schematic showing the typical stages of a hybrid modular approach to SPE. From [7].

cross; in addition to the lack of training data taken in operational scenarios, which creates a large train-test domain gap between the synthetic images which are typically generated for training and real space camera images in deployment.

1.1. Related work

The below summarizes the model which we looked at most closely and worked on refining over the course of the project. For a more extended discussion of prior methods of spacecraft pose estimation, see the survey paper by Pauly et al. [7].

1.1.1 Spacecraft-UDA

Spacecraft-UDA [8, 9], which ranked second in the 2021 Kelvins Pose Estimation Challenge (SPEC2021) [4], is an algorithm for unsupervised domain adaptation (when ground-truth labels are not available for the new domain) with robust pseudo-labeling by inter-model consensus. The model incorporates 3D structure into the spacecraft pose estimation pipeline, via multiple learning losses, to provide robustness against high illumination shifts between domains. It is a keypoint-based approach, whose core idea is to train a model that gets monocular images of the target and returns the expected 2D image location of 3D keypoints and then retrieves pose from the 2D-3D correspondences with PnP, with an unsupervised domain adaptation stage done with iterative pseudo-labeling.

The main model architecture (Fig 4) consists of two stacked hourglass networks which retrieve image keypoint locations with two heads to jointly regress keypoints and associated depth. The motivation for this is that (1) regressing both location and depth of keypoints enables defining a set of learning losses which encourage 3D structure keypoint learning, and (2) a stack of two networks allows the model to retrieve predictions at two different stages in the network. Each network of the hourglass receives an image I and returns two heatmaps: a heatmap $\hat{H} \in R^{N \times M \times C}$ representing the expected image keypoint positions, and a heatmap $\hat{D}^{N \times M \times C}$ representing the depth of each keypoint.

During training, the stacked hourglass is supervised by three losses, which are summed to give the final loss $\ell_{\text{final}} = \ell_{\text{2D-2D}} + \gamma_1 \ell_{\text{2D-3D}} + \gamma_2 \ell_{\text{3D-3D}}$:

1. 2D-2D: consisting of the mean-squared error (MSE) between the estimated heatmap $\hat{H}$ and a ground-truth

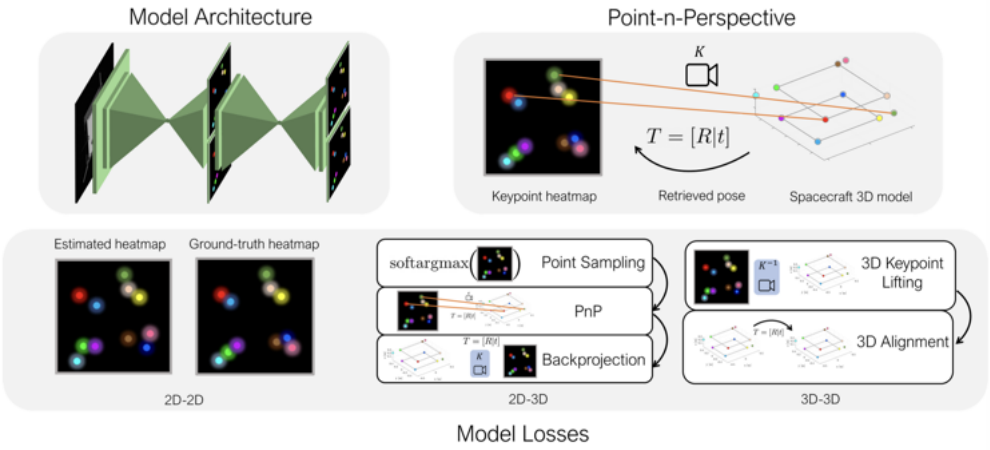

Figure 4. Unsupervised domain adaptation (UDA) model. Figure from [9].

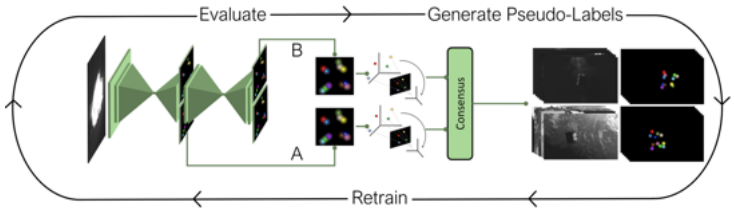

Figure 5. Pseudolabeling procedure for unsupervised domain adaptation (UDA) model. Figure from [9].

heatmap H generated with the ground-truth pose:

$$\ell_{\text{2D-2D}} = \frac{1}{2NMC} \sum_{j=1}^{2} \|\hat{H}^j - \beta H^j\|_F^2 \qquad (1)$$

where $j = 1, 2$ is the index of the stack and $\beta = 100$ controls the gain of the ground-truth Gaussian heatmap to encourage larger loss around keypoint locations.

2. 2D-3D: the difference between the estimated $\hat{p}_i$ and ground-truth 2D keypoint positions p_i compared against each corresponding keypoint position $\dot{p}_i$ generated by projecting each P_i with the pose estimated by PnP.

$$\ell_{\text{2D-3D}} = \frac{1}{N} \sum_{j=1}^{2} \sum_{i=1}^{N} \left(\|p_i^j - \dot{p}_i^j\|_2^2 + \|\hat{p}_i^j - \dot{p}_i^j\|_2^2 \right)$$

3. 3D-3D: the difference between the ground-truth pose T and the pose generated by aligning the 3D estimated keypoints $\bar{P}_i$ with the spacecraft model 3D keypoints P_i, where the bar accent indicates predictions resulting from the 3D alignment.

$$\ell_{\text{3D-3D}} = \frac{1}{24} \sum_{j=1}^{2} \|\bar{T}^j - T^j\|_F^2$$

The unsupervised domain adaptation portion is done by pseudo-labeling with consensus and self-iterative training (Fig. 5). Given the model trained on a source domain, the goal is to generate a set of high-confidence pose predictions $\dot{T}$ on the target domain such that the difference between the true and the estimated labels is small; once this set of pseudo-labels is generated, these are used to fine-tune the model over the new domain. This process is run until a measure for convergence is reached; i.e. until RANSAC converges, which is controlled by (1) the confidence parameter and (2) the reprojection error. The UDA algorithm requires a measurement of consensus to promote pseudo-labeling robustness, which is aided by the hourglass network structure.

2. Methods

2.1. Novel contribution 1: Adaptive wing loss

The base UDA solution considers the pixel-wise mean-squared error (MSE) between the ground-truth H, generated by putting 2D Gaussians of standard deviation = 7px on each projected ground-truth keypoint position, and each predicted heatmap $\hat{H}$, as the 2D-2D loss learning objective (Eq. 1). The issues with MSE for heatmap regression identified by [13] are that (i) MSE is not sensitive to small errors, which is detrimental to correctly locating the mode of the Gaussian distribution, particularly for fore-

ground pixels; and (ii) during training all pixels have the same loss function and equal weights, but background pixels absolutely dominate foreground pixels on a heatmap, which means that models trained with MSE loss incorrectly regress background pixels as foreground pixels and therefore tend to predict a blurry and dilated heatmap with low intensity on foreground pixels when compared to ground truth, particularly for difficult cases such as occluded or unusually-illuminated landmarks.

Furthermore, in the specific context of the SPE problem the MSE formulation treats each keypoint independently, which discourages the network from learning relationships. The UDA authors deal with this by adding the two other losses, but in our project we directly improve on the 2D-2D MSE loss by replacing with an Adaptive Wing (AWing, AW) loss [13]:

$$\text{AWing}(y, \hat{y}) = \begin{cases} \omega \ln\left(1 + \left|\frac{y-\hat{y}}{\epsilon}\right|^{\alpha-y}\right) & \text{if } |y-\hat{y}| < \theta, \\ A|y-\hat{y}| - C & \text{otherwise.} \end{cases} \tag{2}$$

Here y and $\hat{y}$ are the pixel values on the ground-truth and predicted heatmaps respectively, ω, θ, ϵ and α are positive quantities, $A = \omega(1/(1+(\theta/\epsilon)^{\alpha-y}))(\alpha-y)((\theta/\epsilon)^{(\alpha-y-1)})(1/\epsilon)$ and $C = (\theta A - \omega \ln(1+(\theta/\epsilon)^{\alpha-y}))$ are defined to make the loss continuous and smooth at $|y-\hat{y}| = \theta$. θ is a threshold used to switch between linear and nonlinear parts of the function, while ω and ϵ are used to tune the influence on small errors. The exponential term $\alpha - y$ adapts the shape of the loss function to y and makes it smooth at zero; when y pixels have values close to 1 the term is slightly larger than 1 and the nonlinear part behaves like Wing loss, and when y is small as expected for background pixels the gradient behaves more similarly to MSE loss, with a smooth transition in between.

In this way the AWing loss, originally intended to improve facial landmark localization tasks as well as human pose estimation, can adapt its curvature to different types of ground-truth heatmap pixels, penalizing more on foreground pixels and less on background pixels. This reduces small errors on foreground pixels for accurate landmark localization while tolerating small errors on background pixels for better convergence. We also use the Weighted Loss Map described in [13], which assigns high weights to foreground and difficult background pixels to help training process focus more on pixels that are crucial to landmark localization.

2.2. Novel contribution 2: Cascaded Pyramid Network

We further tried replacing the two-stacked hourglass networks of UDA with a Cascaded Pyramid Network (CPN) in order to improve performance on occluded keypoints, invisible keypoints and complex backgrounds. CPN [2] (Figs. 6 and 7), which was originally designed for multi-human pose estimation to replace typical eight-stacked hourglass networks, uses a ResNet backbone, then cascades (1) a GlobalNet feature pyramid network which roughly localizes keypoints, and (2) RefineNet to explicitly handle occluded and otherwise "hard" keypoints by integrating all levels of feature representations from the GlobalNet pyramid features together. In a process referred to as *online hard keypoints mining* (OHKM), RefineNet selects the hard keypoints online based on the training loss and in the backward pass backpropagates the gradients from those selected keypoints only. Unlike refinement strategies like stacked hourglass, RefineNet upsamples and concatenates all pyramid features at the end rather than simply using upsampled features in order to get contextual information. Here we repurpose the CPN feature pyramid as a single stage in our multi-stage design and do OHKM with all three losses from UDA, including our substituted AWing loss.

3. Experiment

3.1. Experiment design

3.1.1 Dataset

Deep learning approaches to SPE still heavily rely on annotated data which are difficult to obtain, hence a continued reliance on synthetic datasets[1]. All models were trained on the SPEED+ dataset [5] by the Stanford Space Rendezvous Laboratory (SLAB), which contains both synthetically generated images and hardware-in-the-loop images of a half-scale mock-up of the Tango satellite whose 11 keypoints are known. There are 60k images in the `synthetic` dataset which is used for training and validation. The testing dataset consists of 2.8k `sunlamp` images which feature strong illumination and reflections against a black background, and 6.8k `lightbox` images with softer illumination but increased noise, as well as a simulated Earth background in a subset of the images.

3.2. Experimental results

3.2.1 Quantitative results

Our results are summarized in Table 1. For each model we compute the translation error score S_t, the Euclidean distance between the estimated translation and ground-truth translation, the rotation error S_r which is the rotation angle between estimated and ground-truth quaternions, and then the final pose score S which combines the two errors.

In addition to a basic SIFT + RANSAC optimization method and PoseResNet, we compare against SPNv2 [3], a recent state-of-the-art end-to-end SPE algorithm by SLAB

[1] Even with synthetic data generation, spacecraft pose datasets typically contain about $10^4 - 10^5$ images, much smaller than datasets such as YCB for example.

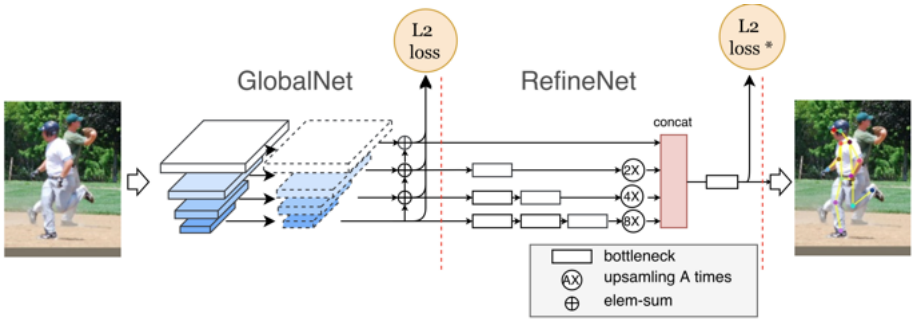

Figure 6. Cascaded Pyramid Network (CPN). "L2 loss*" refers to the L2 loss with online hard keypoints mining. Figure from [2].

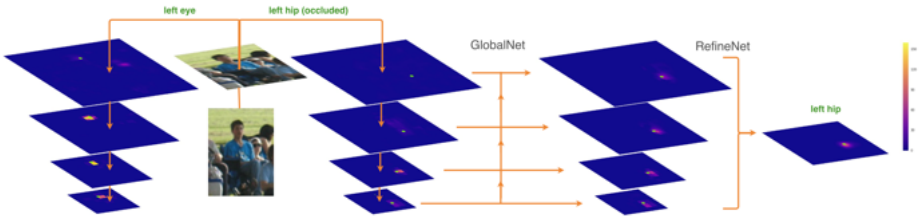

Figure 7. CPN's output heatmaps from different features, where green dots indicate the ground-truth location of keypoints. Shallow features have the high spatial resolution for localization but low semantic information for recognition, while deep feature layers have more semantic information but low spatial resolution due to strided convolution and pooling. Figure from [2].

which builds on prior SLAB work [6, 11], and does robust multi-task learning and online domain refinement by training a model which jointly optimizes multiple prediction heads using a shared EfficientDet [12] feature encoder (EfficientNet backbone with BiFPN to fuse multiple-scale features). The results shown in the table are from offline training of SPNv2 on SPEED+ with $\phi = 3$, with no texture randomization augmentation, trained over 20 epochs; SPNv2 gives two different sets of translation and rotation scores for the heatmap and EfficientPose heads so we just report the final pose loss. The CPN modification ended up being very computationally intensive (due to the need to calculate 5 gradients) but the pose score after 3 epochs was around 2.76.

Model	**Epochs**	S_t	S_r	S
Basic SIFT+RANSAC	-	-	-	-
SPNv2	20	-	-	0.2313
ResNet18 (PoseResNet)	10	1.342	1.2	2.542
Hourglass (Huber/MSE)	10	0.7890	0.3513	1.1403
Hourglass + AWing loss	10	0.7217	0.3618	1.0853

Table 1. Translation S_t, rotation S_r, and final pose error S scores for the modifications we tried relative to several benchmarks (SIFT + RANSAC, PoseResNet, unmodified UDA with stacked hourglasses and MSE loss, SPNv2).

3.2.2 Qualitative results

Figure 8 shows a few examples of the how the stacked-hourglass UDA model with Adaptive Wing loss improves as it learns to predict the keypoints and depth. Figure 9 gives a few examples of the validation results obtained from the model with AWing loss.

4. Conclusions and future work

In this project, we focused our efforts on improving keypoint localization in the spacecraft pose estimation problem, which is a nontrivial hurdle in the hybrid modular approaches typically used to treat the problem. The two improvements we implemented were:

1. Replacing the 2D-2D MSE loss used to supervise the UDA model with an Adaptive Wing (AWing) loss in conjunction with the Weighted Loss Map defined in [13] to improve detection of foreground vs background keypoints;

2. using a Cascaded Pyramid Network (CPN) to first roughly localize keypoints with a GlobalNet and then explicitly handle the "hard" (occluded) keypoints with a RefineNet, which integrates all levels of feature representations from the GlobalNet pyramid features together with an online hard keypoint mining loss (OHKM).

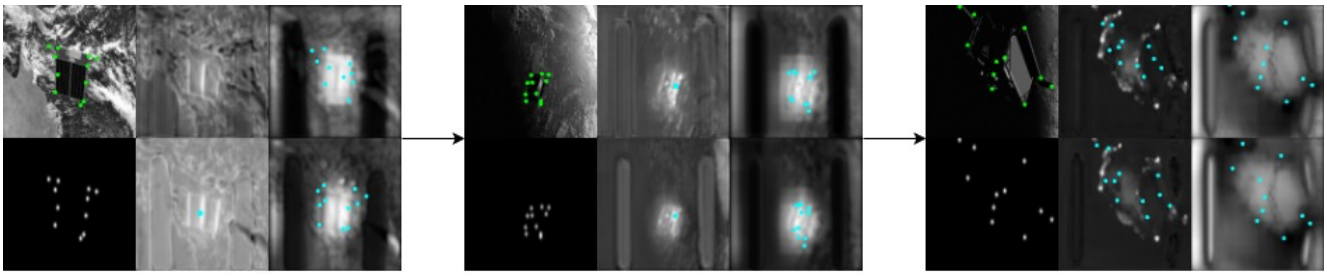

Figure 8. Visualizing the improvement of the two-stacked hourglass model with the AWing loss modification over the course of training. The two rows indicate the stack/pyramid level. The leftmost column of each set of images for each step shows the ground-truth heatmap of the keypoints, the middle column is the keypoint heatmap (PnP) prediction, and the rightmost column is the depth keypoint prediction.

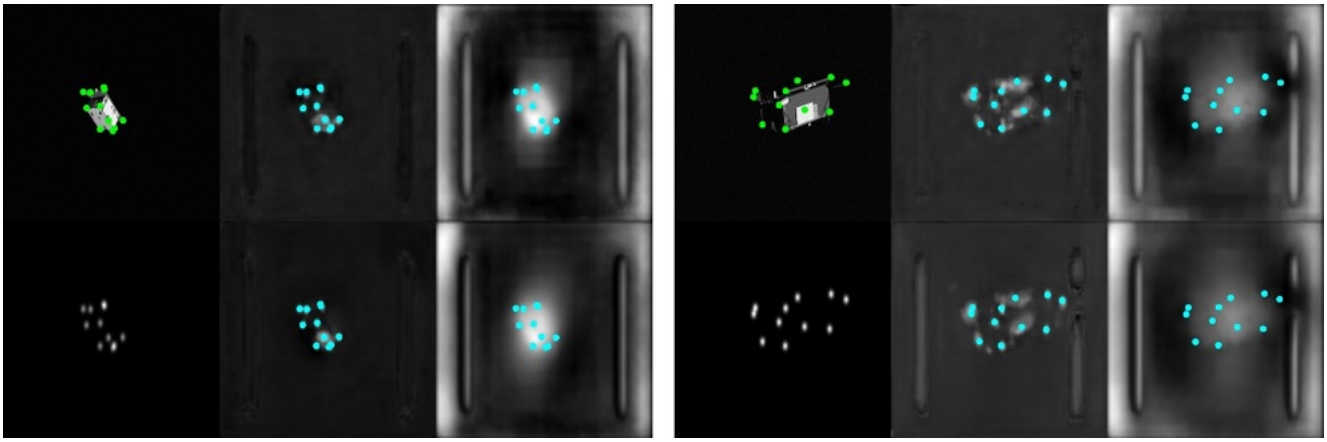

Figure 9. Selected validation results where the rows and columns represent the same information as in the above (Fig. 8).

4.1. Next steps

4.1.1 Heatmap distribution matching

Other 2D-2D losses are possible and could potentially improve upon our keypoint localization results using AWing loss. A further idea we had was to do heatmap distribution matching [10], using the Earth mover distance (EMD) to construct a loss function in order to improve the keypoint localization. The EMD, also known as the Wasserstein metric, measures the difference between two probability distributions as the optimal cost needed to transport the mass from one distribution to another. The regularized Earth mover's distance and its associated loss formulated by [10] is:

$$E_C^{\text{reg}}(S, D) = \langle C, p^{\text{reg}} \rangle \text{ where } p^{\text{reg}} = \arg\min_{p \in P} \left[\langle C, p \rangle - \frac{1}{\lambda} h(p) \right]$$

$$L_{\text{matching}} = \sum_{k=1}^{K} L_k \text{ where } L_k = E_{C^k}^{\text{reg}}(S^k, D^k)$$

which the paper authors show to perform well for human pose estimation for the COCO and MPII datasets.

4.2. Future challenges

The SPEED+ dataset only contains one spacecraft, so there is no need for models to generalize to unknown instances; the most recent (spring 2024) iteration of the Pose Bowl challenge[2] requires generalizing to unknown instances of the object class. This is challenging due to a lack of datasets which contain diverse spacecraft classes *and* the secondary annotations necessary for most current (hybrid modular) approaches, such as bounding boxes, keypoints, segmentation masks, and/or ellipse heatmap annotations. The next challenge is to use 6D pose estimation methods for unseen objects to develop a spacecraft-agnostic pose estimator.

[2] https://www.drivendata.org/competitions/261/spacecraft-pose-estimation/page/837/

References

[1] Bo Chen, Jiewei Cao, Alvaro Parra, and Tat-Jun Chin. Satellite pose estimation with deep landmark regression and nonlinear pose refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1

[2] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7103–7112, 2018. 4, 5

[3] Tae Ha Park and Simone D'Amico. Robust multi-task learning and online refinement for spacecraft pose estimation across domain gap. *Advances in Space Research*, 73(11):5726–5740, 2024. 4

[4] Tae Ha Park, Marcus Märtens, Mohsi Jawaid, Zi Wang, Bo Chen, Tat-Jun Chin, Dario Izzo, and Simone D'Amico. Satellite Pose Estimation Competition 2021: Results and analyses. *Acta Astronautica*, 204:640–665, 2023. 2

[5] Tae Ha Park, Marcus Märtens, Gurvan Lecuyer, Dario Izzo, and Simone D'Amico. SPEED+: Next-generation dataset for spacecraft pose estimation across domain gap. In *2022 IEEE Aerospace Conference (AERO)*, pages 1–15. IEEE, 2022. 4

[6] Tae Ha Park, Sumant Sharma, and Simone D'Amico. Towards robust learning-based pose estimation of noncooperative spacecraft. *arXiv preprint arXiv:1909.00392*, 2019. 5

[7] Leo Pauly, Wassim Rharbaoui, Carl Shneider, Arunkumar Rathinam, Vincent Gaudillière, and Djamila Aouada. A survey on deep learning-based monocular spacecraft pose estimation: Current state, limitations and prospects. *Acta Astronautica*, 2023. 1, 2

[8] Juan Ignacio Bravo Pérez-Villar, Álvaro García-Martín, and Jesús Bescós. Spacecraft pose estimation based on unsupervised domain adaptation and on a 3D-guided loss combination. In *European Conference on Computer Vision*, pages 37–52. Springer, 2022. 2

[9] Juan Ignacio Bravo Pérez-Villar, Álvaro García-Martín, Jesús Bescós, and Marcos Escudero-Viñolo. Spacecraft Pose Estimation: Robust 2D and 3D-Structural Losses and Unsupervised Domain Adaptation by Inter-Model Consensus. *IEEE Transactions on Aerospace and Electronic Systems*, 2023. 1, 2, 3

[10] Haoxuan Qu, Li Xu, Yujun Cai, Lin Geng Foo, and Jun Liu. Heatmap distribution matching for human pose estimation. *Advances in Neural Information Processing Systems*, 35:24327–24339, 2022. 6

[11] Sumant Sharma, Connor Beierle, and Simone D'Amico. Pose estimation for non-cooperative spacecraft rendezvous using convolutional neural networks. In *2018 IEEE Aerospace Conference*, pages 1–12. IEEE, 2018. 5

[12] Mingxing Tan, Ruoming Pang, and Quoc V Le. EfficientDet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10781–10790, 2020. 5

[13] Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *Proceedings of the IEEE/CVF International Conference on Computer vision*, pages 6971–6981, 2019. 3, 4, 5