

TOWARDS ROBUST LEARNING-BASED POSE ESTIMATION OF NONCOOPERATIVE SPACECRAFT

Tae Ha Park*, Sumant Sharma*, Simone D'Amico†

This work presents a novel Convolutional Neural Network (CNN) architecture and a training procedure to enable robust and accurate pose estimation of a noncooperative spacecraft. First, a new CNN architecture is introduced that has scored a fourth place in the recent Pose Estimation Challenge hosted by Stanford's Space Rendezvous Laboratory (SLAB) and the Advanced Concepts Team (ACT) of the European Space Agency (ESA). The proposed architecture first detects the object by regressing a 2D bounding box, then a separate network regresses the 2D locations of the known surface keypoints from an image of the target cropped around the detected Region-of-Interest (RoI). In a single-image pose estimation problem, the extracted 2D keypoints can be used in conjunction with corresponding 3D model coordinates to compute relative pose via the Perspective-n-Point (PnP) problem. These keypoint locations have known correspondences to those in the 3D model, since the CNN is trained to predict the corners in a pre-defined order, allowing for bypassing the computationally expensive feature matching processes. The proposed architecture also has significantly fewer parameters than conventional deep networks, allowing real-time inference on a desktop CPU. This work also introduces and explores the texture randomization to train a CNN for spaceborne applications. Specifically, Neural Style Transfer (NST) is applied to randomize the texture of the spacecraft in synthetically rendered images. It is shown that using the texture-randomized images of spacecraft for training improves the network's performance on spaceborne images without exposure to them during training. It is also shown that when using the texture-randomized spacecraft images during training, regressing 3D bounding box corners leads to better performance on spaceborne images than regressing surface keypoints, as NST inevitably distorts the spacecraft's geometric features to which the surface keypoints have closer relation.

INTRODUCTION

The ability to accurately determine and track the pose (i.e., the relative position and attitude) of a noncooperative client spacecraft with minimal hardware is an enabling technology for current and future on-orbit servicing and debris removal missions, such as the RemoveDEBRIS mission by Surrey Space Centre,¹ the Phoenix program by DARPA,² the Restore-L mission by NASA,³ and GEO servicing programs proposed by Infinite Orbits[‡], Effective Space[§], and many other startup companies. In particular, performing on-board pose estimation is key to the real-time generation of the approach trajectory and control update. The use of a single monocular camera to perform pose estimation is especially attractive due to the low power and mass requirements posed by small spacecraft such as CubeSats. Previous approaches to monocular-based pose estimation⁴⁻⁶ employ

*Ph.D. Candidate, Department of Aeronautics & Astronautics, Stanford University, Stanford, CA 94305

†Assistant Professor, Department of Aeronautics & Astronautics, Stanford University, Stanford, CA 94305

‡www.infiniteorbits.io

§<https://www.effective.space/>

image processing techniques to detect relevant features from a 2D image, which are then matched with features of a known 3D model of the client spacecraft in order to extract relative attitude and position information.⁷ However, these approaches are known to suffer from a lack of robustness due to low signal-to-ratio, extreme illumination conditions, and dynamic Earth background in space imagery. Moreover, these approaches are computationally demanding during pose initialization due to a large search space in determining the feature correspondences between the 2D image and the 3D model.

On the other hand, recent advances in terrestrial computer vision applications of object pose estimation incorporate deep learning algorithms.^{8–21} Instead of relying on explicit, hand-engineered features to compute the relative pose, these algorithms based on deep Convolutional Neural Networks (CNN) are trained to learn the nonlinear mapping between the input images and the output labels, often six-dimensional (6D) pose space or other intermediate information to compute the relative pose. For example, PoseCNN¹⁵ directly regresses relative attitude expressed as a unit quaternion and relative position via separate CNN branches, whereas the network of Tekin et al.¹⁸ modifies the YOLOv2 object detection network²² to regress the 2D corner locations of the 3D bounding box around the object. The detected 2D corner locations can be used in conjunction with corresponding 3D model coordinates to solve the Perspective-n-Point (PnP) problem²³ and extract the full 6D pose. Similarly, KeyPoint Detector (KPD)²⁰ first uses YOLOv3²⁴ to localize the objects, then uses ResNet-101²⁵ to predict the locations of SIFT features²⁶ which can be used in the PnP problem to compute the 6D pose. Recently, PVNet²¹ architecture is proposed to regress the pixel-wise unit vectors which are then used to vote for the location of the keypoints in a way similar to the Random Sample Consensus (RANSAC)²⁷ algorithm. The RANSAC-based voting scheme allows improved prediction accuracy on occluded and truncated objects. PVNet achieves significantly improved performance on LINEMOD and OccludedLINEMOD benchmark datasets.^{28,29}

Not surprisingly, several authors have recently proposed to apply deep CNN to spaceborne pose estimation.^{30–32} Notably, the recent work of Sharma and D’Amico introduced a CNN-based Spacecraft Pose Network (SPN) with three branches that solves for the pose using state-of-the-art object detection network and the Gauss-Newton algorithms.³² The same work also introduced the Spacecraft Pose Estimation Dataset (SPEED) benchmark that contains 16,000 images consisting of synthetic and real camera images of a mock-up of the Tango spacecraft from the PRISMA mission.^{4,33} The dataset is publicly available for researchers to evaluate and compare the performances of pose estimation algorithms and neural networks. Moreover, the SPEED was used in the recent Satellite Pose Estimation Challenge* organized by the Stanford University’s Space Rendezvous Laboratory (SLAB) and the Advanced Concepts Team (ACT) of the European Space Agency (ESA).

However, there are significant challenges that must be addressed before the application of such deep learning-based pose estimation algorithms in space missions. First, the SPN, trained and tested on SPEED, has shown to perform relatively poorly when the spacecraft appears too large or too small in the image.³² Its object detection mechanism also lacked robustness when the spacecraft is occluded due to eclipse. Most importantly, neural networks are known to lack robustness to data distributions different from the one used during training, and it must be verified that these algorithms can meet the accuracy requirements on spaceborne imagery even when trained solely on synthetically generated images. This is especially challenging since spaceborne imagery can contain texture and surface illumination properties and other unmodeled camera artifacts that cannot be perfectly replicated in synthetic imagery. Since spaceborne images are expensive to acquire, the

*<https://kelvins.esa.int/satellite-pose-estimation-challenge/>

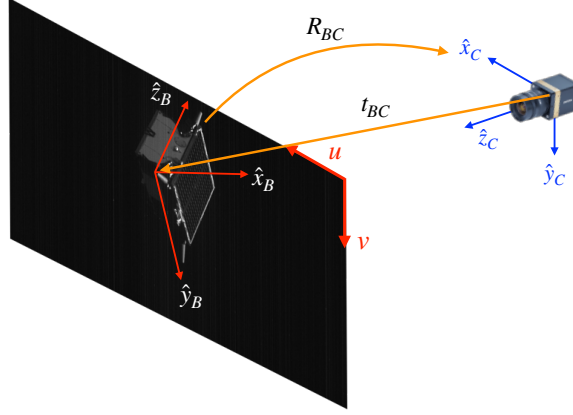


Figure 1. Definition of the body reference frame (B), camera reference frame (C), relative position (t_{BC}), and relative attitude (R_{BC}).

CNN must be able to address this issue with minimal or no access to the properties of spaceborne imagery.

This work makes two contributions to address the aforementioned challenges. The primary contribution of this work is a novel method to enable an efficient learning-based pose determination. Similar to SPN, the problem of pose estimation is decoupled into object detection and pose estimation networks. However, the pose estimation is performed by regressing the 2D locations of the spacecraft's surface keypoints then solving the Perspective-n-Point (PnP) problem. The extracted keypoints have known correspondences to those in the 3D model, since the CNN is trained to predict them in a pre-defined order. This design choice allows for bypassing the computationally expensive feature matching through algorithms such as RANSAC³⁴ and directly use publicly available PnP solvers only once per image.²³ The proposed architecture has scored 4th place in the recent SLAB/ESA Pose Estimation Challenge and is shown to be fast and robust to a variety of illumination conditions and inter-spacecraft separation ranging from 3 to over 30 meters.

The secondary contribution of this work is the study of a novel training procedure that improves the robustness of the CNN to spaceborne imagery when trained solely on synthetic images. Specifically, inspired by the recent work of Geirhos et al., the technique of texture randomization is introduced as a part of the training procedure of the CNN.³⁵ Geirhos et al. suggest that CNN tends to focus on the local texture of the target object, thus randomizing the object texture using the Neural Style Transfer (NST) technique forces the CNN to instead learn the global shape of the object.³⁶ Following their work, a new dataset is generated by applying NST to a custom synthetic dataset that has same pose distribution as SPEED dataset. It is shown that the network exposed to new texture-randomized dataset during training performs better on spaceborne images without having been trained on them.

In the following section, the proposed CNN architecture is explained in detail. The section after that elaborates on the texture randomization procedure and the associated datasets used for training and validation. The section afterward introduces the experiments conducted to evaluate the performance of the proposed CNN and the effect of texture randomization. Finally, the conclusion and the directions for future work are presented.

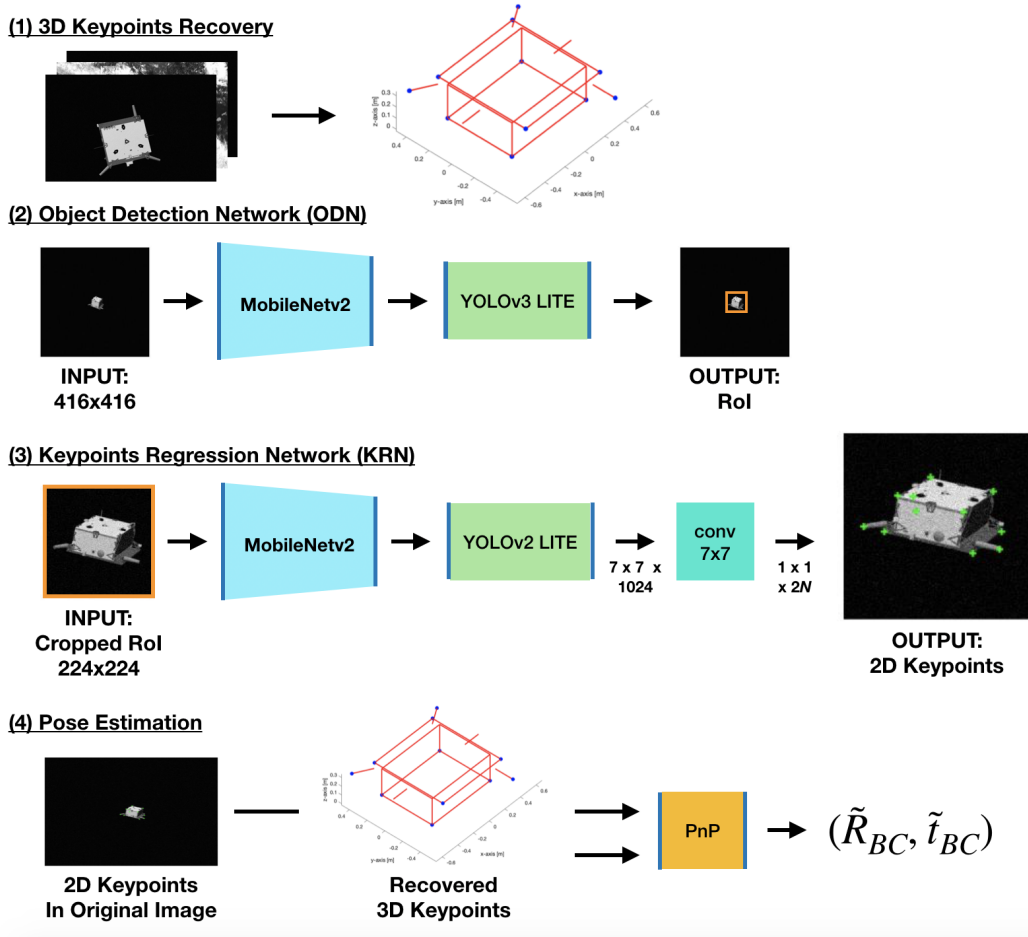


Figure 2. Overall architecture of the proposed CNN.

SINGLE IMAGE POSE ESTIMATION

The general problem statement is to determine the relative attitude and position of the camera frame, \mathcal{C} , with respect to the target's body frame, \mathcal{B} . The relative position is represented by a position vector, t_{BC} , from the origin of \mathcal{C} to the origin of \mathcal{B} . Similarly, the relative attitude is represented by a rotation matrix, R_{BC} , which aligns the reference frame \mathcal{B} with \mathcal{C} . Figure 1 graphically illustrates these reference frames and variables.

The overall pipeline of the single image pose estimation architecture developed in this work is visualized in Figure 2 in four steps.

1. First, the 3D model coordinates of 11 keypoints are selected from the available wireframe model of the Tango spacecraft. If the model is not available, the 3D model coordinates of the selected keypoints are recovered from a set of training images and associated pose labels. Figure 3 visualizes the keypoints selected for this architecture, which geometrically correspond to four corners of the bottom plate, four corners of the top plate (i.e. solar panel), and three tips of the antennae.
2. Second, the Object Detection Network (ODN) detects a 2D bounding box around the space-

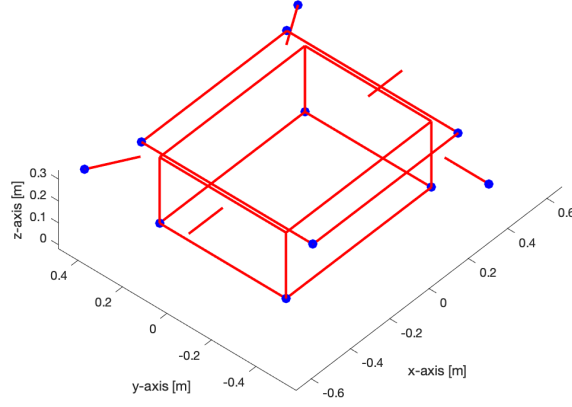


Figure 3. 11 keypoints used in the proposed architecture visualized on a wireframe model of Tango spacecraft³² (blue dots)

craft from the image resized to 416×416 . The 2D bounding box labels are obtained by projecting the 3D keypoints onto an image plane using the provided ground-truth poses, then taking maximum and minimum coordinates in x and y directions.

3. Third, the detected 2D bounding box is used to crop the Region-of-Interest (RoI) from the original image, which is resized to 224×224 and fed into the Keypoints Regression Network (KRN). The KRN returns $1 \times 2N$ vector encoding the 2D locations of N keypoints.
4. Lastly, the extracted 2D keypoints are measured in the context of the original image. Then, they can be used in solving the PnP problem using an off-the-shelf PnP solver along with the known or recovered 3D model coordinates to compute the full 6D pose.

3D Keypoints Recovery

While this work exploits the available wireframe model of the Tango spacecraft, the method of 3D keypoints recovery is introduced for completeness. In order to recover the 3D model coordinates (\mathbf{p}_{3D}) of the aforementioned 11 keypoints, a set of 12 training images is selected in which the Tango spacecraft is well-illuminated and has varying poses. Then, a set of visible keypoints (\mathbf{p}_{2D}) is manually picked. In order to recover \mathbf{p}_{3D} , the following optimization problem is solved,

$$\text{minimize } \sum_j ||s_j \mathbf{p}_{2D,k}^h - \mathbf{K}[\mathbf{R}_j | \mathbf{t}_j] \mathbf{p}_{3D,k}^h||_2 \quad (1)$$

where, for each k -th point, the sum of the reprojection error is minimized over a set of images in which the k -th point is visible. In Eq. (1), a superscript h indicates the point is expressed in homogenous coordinates, \mathbf{K} is a known camera intrinsic matrix, and $(\mathbf{R}_j, \mathbf{t}_j)$ is a known pose associated with the j -th image. The optimization variables in Eq. (1) are the 3D model coordinates, $\mathbf{p}_{3D,k}$, and scaling factors, s_j , associated with the projection onto the image plane in each input image. Since Eq. (1) is a convex objective function of its optimization variables, the solutions, $(\mathbf{p}_{3D,k}^h, s_j) \forall j = 1, \dots, 11$, are obtained using the CVX solver.^{37,38}

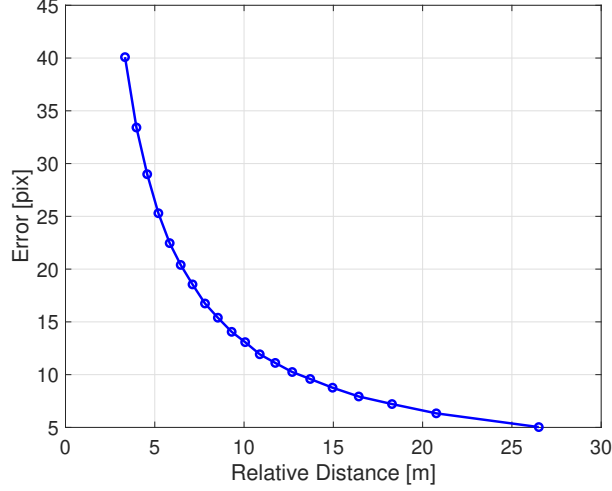


Figure 4. Average of reprojection error of recovered 3D keypoints plotted against mean relative distance, $\|t_{BC}\|_2$, for the SPEED synthetic training images.

Overall, the reconstructed 3D keypoint coordinates have an average distance error of 5.7 mm compared to those in the wireframe model, with the maximum error around 9.0 mm. Figure 4 plots the reprojection error of the recovered keypoints against the ground-truth keypoints from the wireframe model. While the maximum average error is around 40 pixels, the CNN trained with labels from recovered keypoints implicitly learns the offsets from the ground-truth coordinates.

Object Detection Network (ODN)

The ODN pipeline closely follows the structure of the state-of-the-art detection network, YOLOv3.²⁴ It takes an input of 416×416 and performs detection at three different stages. The original backbone of Darknet-53 and extra convolutional layers are replaced with MobileNetv2³⁹ and depth-wise separable convolution operations,⁴⁰ respectively, to drastically reduce the number of parameters in the network. Specifically, depth-wise separable convolution breaks up the conventional convolution into depth-wise and point-wise convolution operations. As the names suggest, depth-wise convolution applies a single kernel per each channel, compared to conventional convolution that applies kernels over all channels. On the other hand, point-wise convolution applies a 1×1 kernel over all channels and, unlike depth-wise convolution, can be used to output arbitrary number of channels. Figure 5 qualitatively describes different convolution operations. As in MobileNet,⁴⁰ both depth-wise and point-wise convolutions are followed by batch normalization and Rectified Linear Unit (ReLU) activation function layers.

In general, for a convolution unit with kernel size $K \times K$, C_{in} input channels and C_{out} output channels, the number of tunable parameters is given as

$$K \times K \times C_{in} \times C_{out}, \quad (2)$$

whereas depth-wise separable convolution has

$$K \times K \times C_{in} \times 1 + 1 \times 1 \times C_{in} \times C_{out} \quad (3)$$

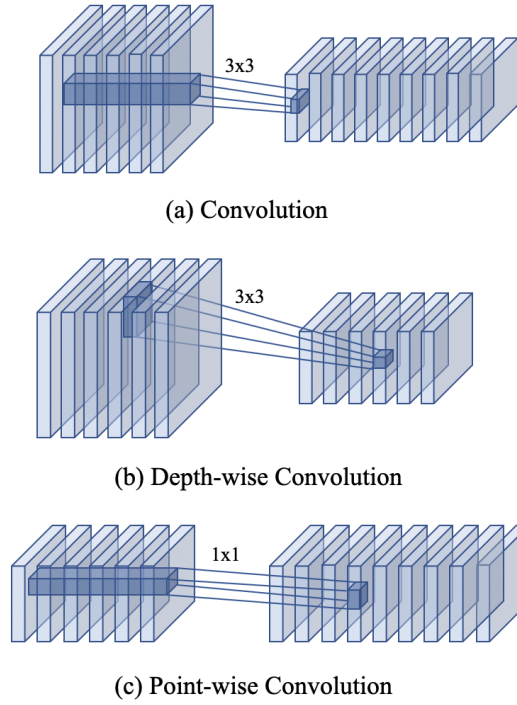


Figure 5. Different convolution operations. In this work, conventional convolution operation (a) is replaced by depth-wise convolution (b) followed by the point-wise convolution (c).

parameters. The factor of reduction in number of parameters is then given as

$$\frac{1}{C_{\text{out}}} + \frac{1}{K \times K}. \quad (4)$$

Given that the most common kernel size in state-of-the-art CNN architectures is 3×3 , simply replacing each convolution with depth-wise separable convolution reduces computation by a factor of 8 or 9.

Since SPEED guarantees the presence of a single, known spacecraft in every image, no classification is performed in the ODN. The direct output of the ODN is an $N \times N \times 5$ tensor, where $N = 13, 26, 52$ in respective prediction stages. The output tensor essentially divides the input image into $N \times N$ grids, each grid predicting $(t_0, t_x, t_y, t_w, t_h)$. These predictions are related to the objectness score, $p(c)$, the location of the bounding box center, (x, y) , and the size of the bounding box, (w, h) , via the following equations,

$$\begin{aligned} p(c) &= \sigma(t_0) \\ x &= \sigma(t_x) + g_x \\ y &= \sigma(t_y) + g_y \\ w &= p_w e^{t_w} \\ h &= p_h e^{t_h} \end{aligned} \quad (5)$$

where $\sigma(x)$ is a sigmoid function, (g_x, g_y) is the location of each grid, and (p_w, p_h) is the size of each anchor box. Similar to YOLOv3, a total of nine anchor boxes are pre-defined using k-means

clustering, and three are used for prediction in each stage. The ground-truth objectness score is assigned 1 to a grid containing the object and the best-matching anchor. Since there is only one object, the prediction is made from the grid with the highest objectness score during the inference without non-max suppression. The readers are encouraged to refer to a series of publications on YOLO architecture for more details on the implementation.^{22,24,41}

Compared to the original YOLOv3, the loss function is modified to better maximize the Intersection-over-Union (IoU) metric defined as

$$\text{IoU} = \frac{I}{U} = \frac{\text{Area of Intersection}}{\text{Area of Union}}. \quad (6)$$

Specifically, the Mean-Squared Error (MSE) loss of the bounding box parameters is replaced by the Generalized Intersection-over-Union (GIoU) loss defined as⁴²

$$\mathcal{L}_{\text{GIoU}} = 1 - \text{GIoU}, \text{ where } \text{GIoU} = \frac{I}{U} - \frac{A^C - U}{A^C}, \quad (7)$$

where A^C is an area of the smallest box enclosing both predicted and ground-truth bounding boxes. The formulation of GIoU loss ensures that the gradient is bigger as the separation between bounding boxes becomes larger even when there is no overlap (i.e. $\text{IoU} = 0$). The overall loss, excluding the classification loss, is then

$$\lambda_{\text{GIoU}} \mathcal{L}_{\text{GIoU}} + \lambda_{\text{conf}} \mathcal{L}_{\text{conf}} \quad (8)$$

where λ_{GIoU} and λ_{conf} are weighting factors, and $\mathcal{L}_{\text{conf}}$ is a sum of the Binary Cross-Entropy (BCE) loss between predicted and ground-truth objectness score.

Keypoints Regression Network (KRN)

The input to the KRN is a RoI cropped from the original image using the 2D bounding box detected from the ODN. The motivation behind the cropping is the fact that the SPEED images are large (1920×1200 pixels) compared to the input sizes typically demanded by CNN architectures (e.g. 224×224 for VGG-based networks). Regular resizing from 1920×1200 to 224×224 will blur much of the detailed features that can help make accurate prediction of the keypoint locations, especially if the target appears small due to large inter-spacecraft separation. Therefore, cropping the RoI prior to KRN helps the network make better predictions based on much finer features. In general, this approach works regardless of the image size and makes the architecture robust to different feature resolutions.

The structure of the KRN closely follows the architecture of YOLOv2 but exploits the MobileNet architecture and depth-wise separable convolution operations similar to ODN. It receives the input, which is cropped around the RoI and resized to 224×224 , and outputs a $7 \times 7 \times 1024$ tensor. The output tensor is reduced to a $1 \times 2N$ vector via a convolution with 7×7 kernel to regress the 2D locations of the N keypoints, where $N = 11$ as defined earlier. It is empirically found that dimension reduction using a convolution performs better than using global average pooling. These keypoints are then used to compute the 6D pose estimate using the EPnP algorithm²³ with the selected 3D keypoint coordinates. The loss function of the KRN is simply a sum of MSE between the predicted ($\tilde{\mathbf{k}}$) and ground-truth keypoint (\mathbf{k}) locations, i.e.

$$\mathcal{L}_{\text{KRN}} = \sum_{j=1}^{11} \|\tilde{k}_x^{(j)} - k_x^{(j)}\|_2 + \|\tilde{k}_y^{(j)} - k_y^{(j)}\|_2. \quad (9)$$

TEXTURE RANDOMIZATION

The advantage of many benchmark datasets for various deep learning-based computer vision tasks, such as ImageNet,⁴³ MS COCO,⁴⁴ or LINEMOD,²⁸ is that they comprise images from the real world to which the CNNs are expected to be applied. However, due to the difficulty of acquiring the same amount of spacecraft images with accurately annotated labels, the training dataset for spaceborne CNN inevitably depends heavily on synthetic renderers and laboratory testbeds. Unfortunately, it is extremely difficult to exactly replicate the target spacecraft’s surface properties and lighting conditions encountered throughout a space mission. Since any CNN for space application is likely to be trained mainly on a set of synthetic images, the gap between the properties of synthetic and spaceborne images must be addressed to ensure the CNN’s functionality in space missions. While it is possible to include some spaceborne images during training to improve the CNN’s generalizability, this work specifically considers only the availability of synthetic images during training.

Arguably, one of the most distinct differences between the synthetic and spaceborne images is surface texture. Recently, Geirhos et al. have empirically shown that many state-of-the-art CNN architectures, such as ResNet²⁵ and VGG,⁴⁵ exhibit strong bias toward object’s texture when trained on ImageNet.³⁵ This finding is very counter-intuitive for humans who would, for example, classify an object or an animal based on its global or a set of local shapes, not based on its texture. To demonstrate such behavior, the authors have created Stylized-ImageNet (SIN) dataset by applying the AdaIN Neural Style Transfer (NST) pipeline⁴⁶ to each image in ImageNet with random styles from Kaggle’s *Painter by Numbers* dataset*. The result shows that when the same CNN architectures are trained instead on SIN dataset, the networks not only exceed the performance of those trained on ImageNet, they also show human-level robustness to previously unseen image distortions, such as noise, contrast change, and high- or low-pass filtering.

In this work, a similar approach is adopted to emphasize the significance of the spacecraft texture on the CNN’s performance. First, a synthetic dataset PRISMA12K is created using the same rendering software used in SPEED. PRISMA12K consists of 12,000 synthetic images of the Tango spacecraft with the same pose distribution as SPEED. However, PRISMA12K uses the camera model of the vision-based sensor used on the Mango spacecraft during the PRISMA mission (Table 1).

Table 1. PRISMA camera parameters and values.

Parameter	Description	Value
N_u	Number of horizontal pixels	752
N_v	Number of vertical pixels	580
f_x	Horizontal focal length	0.0200 m
f_v	Vertical focal length	0.0193 m
du	Horizontal pixel length	8.6×10^{-6} m
dv	Vertical pixel length	8.3×10^{-6} m

A second dataset, PRISMA12K-TR, is created by applying a NST pipeline to each image of the PRISMA12K offline to randomize the spacecraft texture. In this work, the pre-trained NST pipeline proposed by Jackson et al. is used.³⁶ Instead of explicitly supplying the style image at each

*<https://www.kaggle.com/c/painter-by-numbers/>

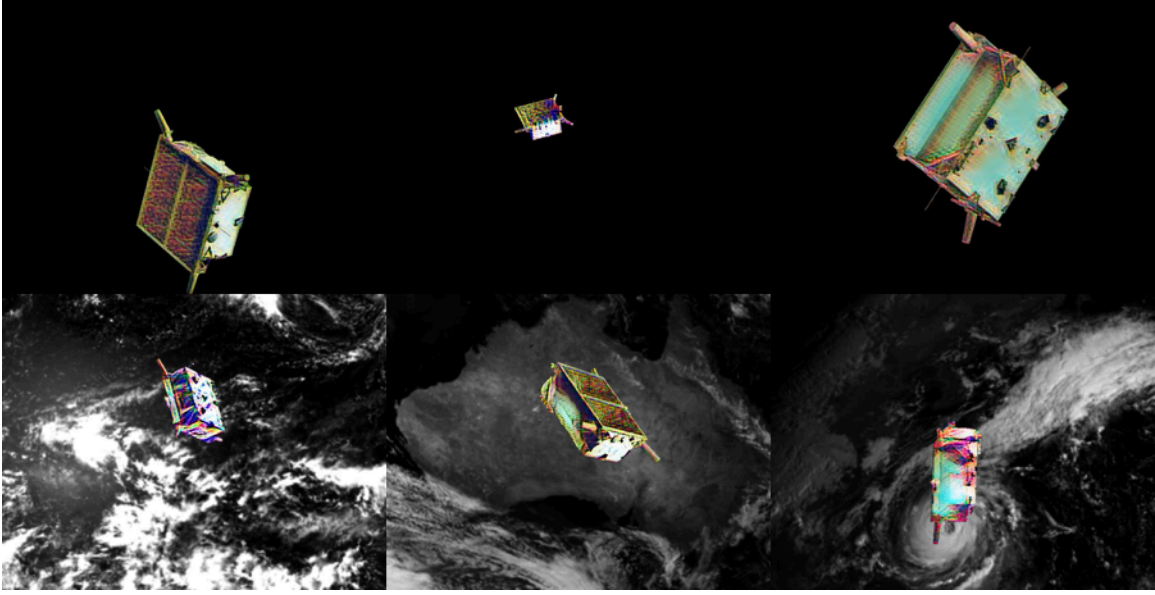


Figure 6. Examples of 6 images from PRISMA12K-TR

inference, this NST pipeline allows for randomly sampling a vector of style embedding $z \in \mathcal{R}^{100}$. Specifically, the style embedding is sampled as

$$z = \alpha \mathcal{N}(\mu, \Sigma) + (1 - \alpha)P(c) \quad (10)$$

where $P(c)$ is the style embedding of the content image, (μ, Σ) are the mean vector and covariance matrix of the style image embeddings pre-trained on ImageNet, and α is the strength of the random normal sampling. In this work, $\alpha = 0.25$ is used to create PRISMA12K-TR. In order to avoid the NST’s blurring effect on the spacecraft’s edges, the style-randomized spacecraft is first isolated from the background using a bitmask then combined with the original background. Figure 6 shows a montage of six such images.

The third dataset is PRISMA25, which consists of 25 spaceborne images captured during the rendezvous phase of the PRISMA mission.³³ The PRISMA25 is used to evaluate the performance of the CNN on a previously unseen spaceborne dataset when trained solely on a mixture of PRISMA12K and PRISMA12K-TR.

EXPERIMENTS

In this section, the procedures and results of two experiments are elaborated. Throughout both experiments, two variants of the keypoint regression network are trained and evaluated. The first variant, noted as KRN-SK, is identical to the KRN introduced in Section 2 and regresses the 2D coordinates of 11 surface keypoints. The second variant, noted as KRN-BB, instead regresses the 2D coordinates of the centroid and eight corners of the 3D bounding box around the object.

The first experiment evaluates the performance of the proposed single image pose estimation architecture, namely ODN and KRN-SK. In order to provide an in-depth analysis, both networks are first trained on 80% of the 12,000 synthetic training images and evaluated on the validation set which comprises the rest of the 20%. The performance of the combined architecture on synthetic

and real test sets is also reported. The second experiment instead trains KRN-SK and KRN-BB using mixtures of PRISMA12K and PRISMA12K-TR using the ground-truth 2D bounding boxes. Both versions of KRN are evaluated on PRISMA25 for comparison and in order to gauge the effect of texture randomization in closing the domain gap between synthetic and spaceborne images. The keypoint labels are generated using the ground-truth wireframe model of the Tango spacecraft unless stated otherwise.

Evaluation Metrics

Throughout this section, four performance metrics are used to evaluate the proposed architecture. For ODN, the mean and median IoU scores are reported as in Eq. (6) to measure the degree of overlap between the predicted and ground-truth 2D bounding boxes. For the combined architecture, mean and median translation and rotation errors are reported as³²

$$E_T = |\tilde{\mathbf{t}}_{BC} - \mathbf{t}_{BC}|, \quad (11)$$

$$E_R = 2 \arccos |\mathbf{q}_{BC} \cdot \tilde{\mathbf{q}}_{BC}| \quad (12)$$

where $(\tilde{\mathbf{q}}_{BC}, \tilde{\mathbf{t}}_{BC})$ are predicted unit quaternion and translation vector aligning the target body frame (\mathcal{B}) and camera frame (\mathcal{C}), and $(\mathbf{q}_{BC}, \mathbf{t}_{BC})$ are ground-truth unit quaternion and translation vector. Lastly, the pose score used in SLAB/ESA Pose Estimation Challenge (henceforth noted SLAB/ESA score) is reported as

$$\text{SLAB/ESA score} = \frac{1}{N} \sum_{i=1}^N \frac{\|\tilde{\mathbf{t}}_{BC}^{(i)} - \mathbf{t}_{BC}^{(i)}\|_2}{\|\mathbf{t}_{BC}^{(i)}\|_2} + E_R^{(i)}. \quad (13)$$

Experiment 1: Single Image Pose Estimation

For single image pose estimation, both ODN and KRN are trained using the RMSprop optimizer⁴⁷ with batch size of 48 and momentum and weight decay set to 0.9 and 5×10^{-5} , respectively, unless stated otherwise. For both networks, the learning rate is initially set to 0.001 and decays exponentially by a factor of 0.98 after every epoch. The networks are implemented with PyTorch v1.1.0 and trained on an NVIDIA GeForce RTX 2080 Ti 12GB GPU for 100 epochs for ODN and 300 for KRN. No real images are used in training to gauge the architecture’s ability to generalize to the datasets from different domains.

Table 2. Parameters and distributions of data augmentation techniques. Brightness, contrast, and Gaussian noise are implemented with 50% chance during training.

Brightness (β)	Contrast (α)	Gaussian Noise	RoI Enlargement Factor [%]	RoI Shifting Factor [%]
$\mathcal{U}(-25, 25)$	$\mathcal{U}(0.5, 2.0)$	$\mathcal{N}(0, 25)$	$\mathcal{U}(0, 50)$	$\mathcal{U}(-10, 10)$

In training both networks, a number of data augmentation techniques is implemented. For both ODN and KRN, the brightness and contrast of the images are randomly changed according to

$$p'(i, j) = \alpha p(i, j) + \beta \quad (14)$$

where $p(i, j) \in [0, 255]$ is the value of a pixel at i^{th} column and j^{th} row of the image. The images are randomly flipped and rotated at 90° intervals, and a random Gaussian noise is also implemented.

Table 3. Performance of the proposed architecture on synthetic validation set.

Metrics	SPEED Synthetic Validation Set
Mean IoU	0.919
Median IoU	0.936
Mean E_T [m]	[0.010, 0.011, 0.210]
Median E_T [m]	[0.007, 0.007, 0.124]
Mean E_R [deg]	3.097
Median E_R [deg]	2.568
SLAB/ESA Score	0.073

For KRN, the ground-truth RoI is first corrected to a square-sized region with a size of $\max(w, h)$, where (w, h) are the width and height of the original RoI. This correction is implemented to ensure the aspect ratio remains the same when resizing the cropped region into 224×224 . Then, the new square RoI is enlarged by a random factor up to 50% of the original size. Afterwards, the enlarged RoI is shifted in horizontal and vertical directions by a random factor up to 10% of the enlarged RoI dimension. This technique has an effect of making the network robust to object translation and misaligned RoI detection. During testing, the detected RoI is similarly converted into a square-sized region and enlarged by a fixed factor of 20% to ensure the cropped region contains the entirety of the spacecraft. The distributions of each augmentation parameter are summarized in Table 2.

Table 3 reports the proposed CNN’s performance on the SPEED synthetic validation dataset. Overall, the ODN excels in detecting the spacecraft from the images with mean IoU of 0.919. The worst IoU score in the validation set is reported as 0.391, indicating that even the worst prediction of the 2D bounding box still has some overlap with the target, mitigating the effect of misaligned RoI on the keypoints regression. The pose solutions of the combined architecture also show improved performance on synthetic validation set compared to that of SPN.³² Specifically, the mean E_T is under 25 cm, while the mean E_R is around 3.1 degrees.

Figure 7 visualizes six cases of successful pose predictions. Overall, the figures demonstrate both ODN and KRN are able to make accurate predictions despite clipping due to proximity, severe shadowing, or large inter-spacecraft separation above 30 meters regardless of the presence of Earth in the background. However, as shown in the cases of the four worst predictions visualized in Figure 8, the CNN is not always immune to large separation or extreme shadowing. Figure 8 demonstrates that the combined networks, despite the “zooming-in” effect from cropping around the RoI, can still fail in accurate keypoints regression when the inter-spacecraft separation is too large because the keypoint features of the spacecraft become indistinguishable from the Earth in the background. In other case, the boundary of the spacecraft’s shape blurs and blends into the Gaussian noise in the background due to the shadowing and large separation. On the other hand, the ODN is able to predict the bounding boxes with high accuracy even in the four worst prediction cases.

Figure 9 plots the average of the translation and rotation errors with respect to the mean ground-truth relative distance. The distribution of the errors exhibits the trend also visible in SPN³² – the position error grows as the spacecraft is farther away, and the rotation error is worse when spacecraft is too close or too far. Specifically, the mean rotation error for the largest inter-spacecraft distance suffers from extreme outliers, such as the case visualized in the top-left figure of Figure 8. However, in general, the mean translation error is under a meter due to the successful performance of ODN

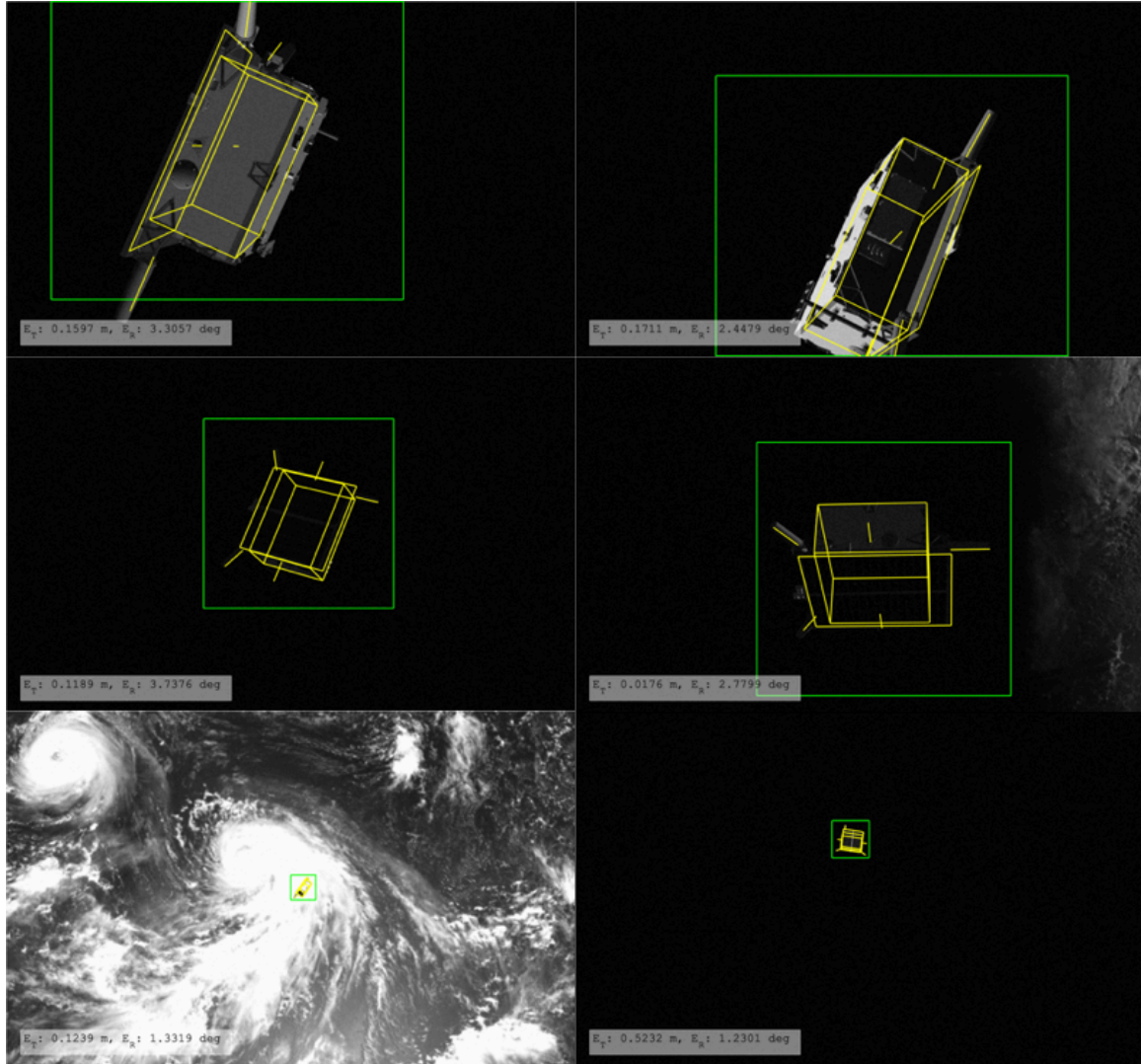


Figure 7. Examples of the predicted 2D bounding boxes and pose solutions of the proposed architecture on the SPEED validation set. The 2D bounding boxes shown are enlarged by 20% from the predicted boxes. In general, the proposed CNN excels in the cases of proximity (top), severe shadowing (middle), and large separation (bottom).

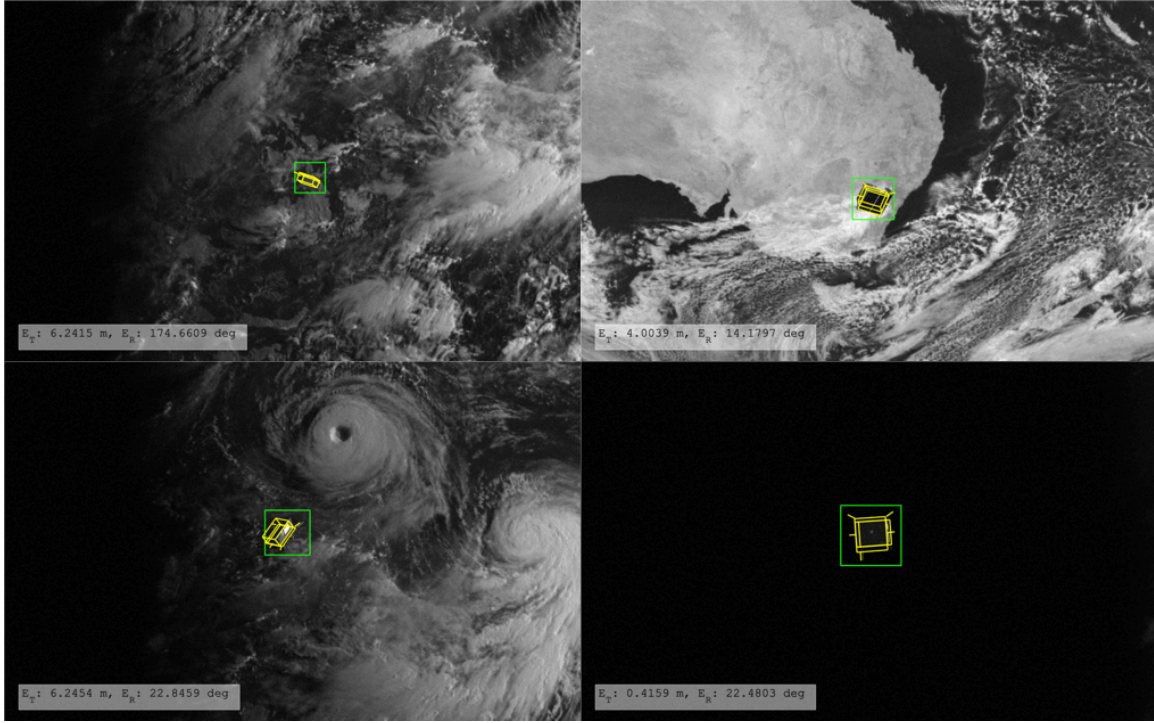


Figure 8. Four worst pose solutions on the SPEED validation set. The 2D bounding boxes shown are enlarged by 20% from the predicted boxes. Even with cropping around the RoI, in some cases the spacecraft features blend into the Earth background (top) or Gaussian noise (bottom-right).

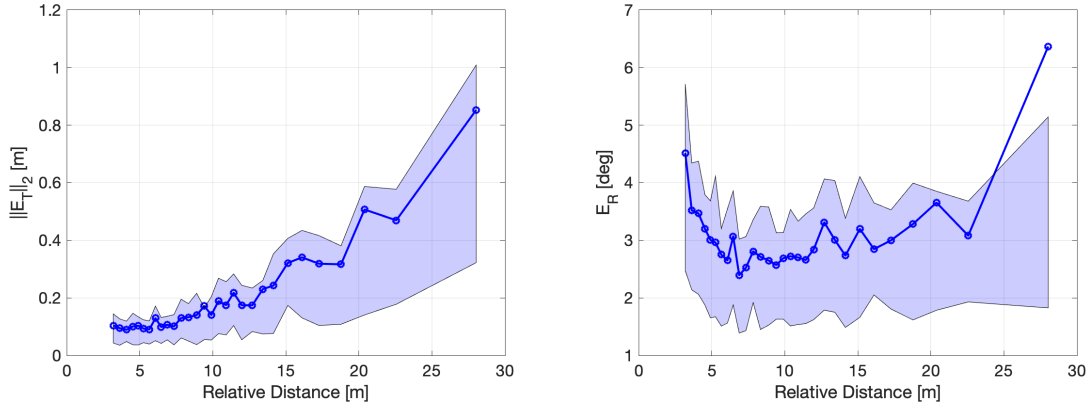


Figure 9. Mean $\|E_T\|_2$ and E_R plotted against mean relative distance, $\|t_{BC}\|_2$, for the SPEED validation set. The shaded region shows 25 and 75 percentile values.

for all range of inter-spacecraft separation, and unlike SPN, the clipping due to proximity does not cause spike in translation error thanks to the random RoI shifting during training.

Table 4 lists the number of parameters and inference time associated with each network in the proposed architecture. Due to the MobileNet architecture and innovative depth-wise convolution operations, the proposed architecture requires less computation despite exploiting the architectures of the state-of-the-art deep networks. For example, the YOLOv2-based KRN only has 5.64 million

Table 4. Size and speed of the proposed architecture. Inference speed only accounts for the forward propagation of the ODN and KRN without any post-processing steps.

Network	Number of parameters [Millions]	Size [MB]	Runtime on GPU [ms]	Runtime on CPU [ms]
ODN	5.53	22.4	7	230
KRN	5.64	22.8	7	32
Total	11.17	45.2	14	262

parameters compared to about 50 million of the YOLOv2. By itself, the KRN runs at 140 Frames Per Second (FPS) on GPU and about 30 FPS on an Intel® Core™ i9-9900K CPU at 3.60GHz for inference. Similar trend can be observed for the YOLOv3-based ODN; however, the inference time on CPU increases dramatically most likely due to the upsampling operations inherent to the YOLOv3 architecture. Overall, the combined architecture runs at about 70 FPS on a GPU and 4 FPS on a CPU. An architecture like MobileNet can potentially pave the way towards the implementation of deep CNN-based algorithms on-board the spacecraft with limited computing resources.

Table 5. Performance of the proposed architecture on SPEED test sets. In this case, the recovered keypoints are used as labels during training.

Metric	SPEED Synthetic Test Set	SPEED Real Test Set
SLAB/ESA Score	0.0626	0.3951
Placement	4th	4th

The SLAB/ESA scores on both synthetic and real *test* sets are also reported in Table 5. In this case, both networks are trained with all 12,000 synthetic training images. It is clear that with a bigger training dataset, the score decreases compared to that reported in Table 3. With the synthetic score of 0.0626, the proposed architecture has scored 4th place in the SLAB/ESA Pose Estimation Challenge. However, because the training only involved the synthetic images, the score on the real test set is about six times worse than that on the synthetic test set.

Experiment 2: Texture Randomization

For texture randomization, only the performances of the KRNs are tested, as the object detection has shown to improve with texture-randomized training images from the literature.³⁶ In this experiment, AdamW optimizer⁴⁸ is used with momentum and weight decay set to 0.9 and 5×10^{-5} , respectively. The learning rate is initially set to 0.0005 and halves after every 50 epochs. Both KRN-SK and KRN-BB are trained for 200 epochs on the same GPU hardware as introduced in Experiment 1.

Both variants of the KRN are trained using the ground-truth RoI that are randomly enlarged and shifted similar to the first experiment. For each input image, the network chooses an image from PRISMA12K-TR over PRISMA12K with probability of p_{TR} . For images from PRISMA12K, the same data augmentation techniques in Table 2 are used, except Gaussian noise is sampled from $\mathcal{N}(0, 10)$. For images from PRISMA12K-TR, random erasing augmentation technique⁴⁹ is applied with 50% probability in order to mimic the shadowing effect due to the eclipse. This is because the NST cancels any illumination effect that was cast on the spacecraft, as seen in Figure 6.

Table 6. Performance of the KRN-BB on PRISMA25 for varying p_{TR} . The average and standard deviation of SLAB/ESA score over 3 tests are reported for the best and last epochs. Bold face numbers represent the best performances.

	$p_{\text{TR}} = 0$	$p_{\text{TR}} = 0.25$	$p_{\text{TR}} = 0.50$	$p_{\text{TR}} = 0.75$
Best	0.927 ± 0.072	0.717 ± 0.276	0.513 ± 0.102	0.849 ± 0.133
Last	1.388 ± 0.494	0.884 ± 0.280	0.943 ± 0.158	1.246 ± 0.209

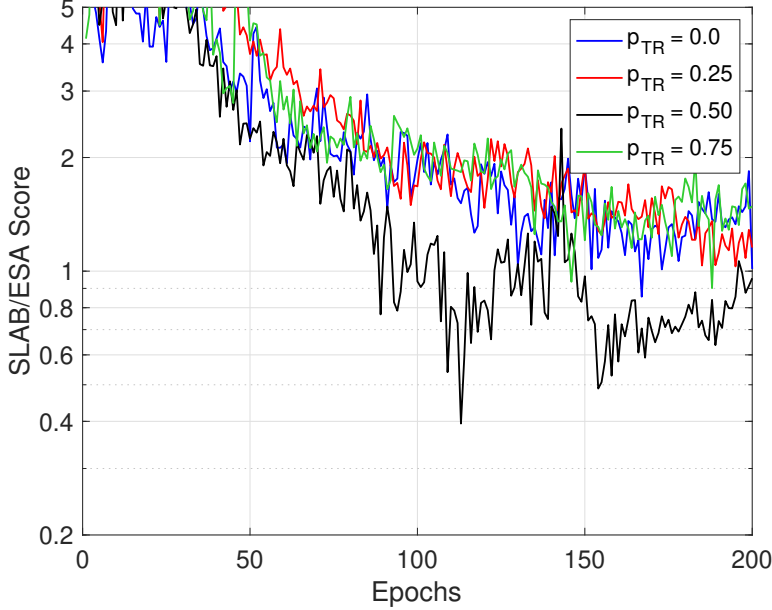


Figure 10. SLAB/ESA scores on PRISMA25 during training.

Table 6 reports the SLAB/ESA scores of the KRN-BB on PRISMA25 with varying p_{TR} . Specifically, the experiments are run three times with different random seeds to check the consistency in training behavior, and the averaged scores are reported using the network after the best-performing epoch (Best) and the last epoch (Last). First, according to the scores at the best-performing epochs, the KRN-BB with $p_{\text{TR}} = 0.5$ consistently achieves the lowest SLAB/ESA score compared to the network with other p_{TR} or the network trained purely on synthetic PRISMA12K (i.e. $p_{\text{TR}} = 0$). However, the SLAB/ESA scores reported after the training is complete gives an impression that there is no visible improvement when the images from PRISMA12K-TR are introduced during training. The reason is that the network’s performance on spaceborne images from PRISMA25 becomes very volatile as the training nears the end, as visible in Figure 10. However, despite the volatility, it is obvious from Figure 10 that the training with $p_{\text{TR}} = 0.5$ outperforms the other study cases in general throughout the training.

One strong candidate for the cause of volatility on PRISMA25 is the fact that the texture randomization via NST inevitably disrupts the local shapes and geometric features of the spacecraft. For example, Figure 11 visualizes three cases in which the asymmetric parts of the spacecraft, such as antennae and features on the spacecraft body, are rendered indistinguishable due to the NST pipeline. Then, it is not surprising that the evaluation on PRISMA25 becomes volatile, since the

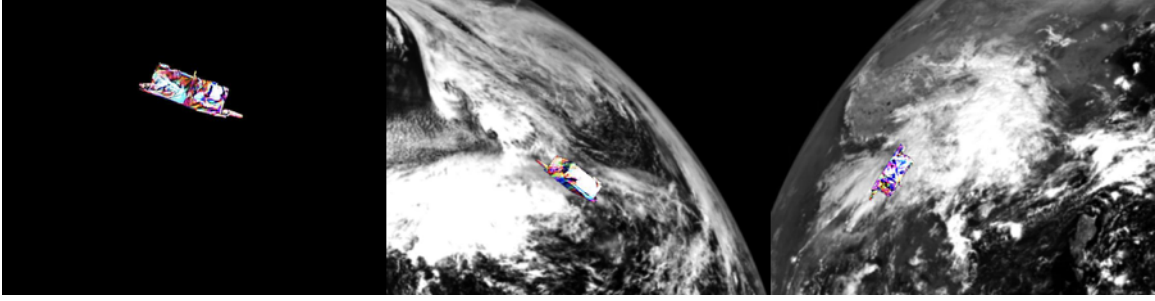


Figure 11. Examples of bad texture randomization.

Table 7. Performance of both KRN on PRISMA25 for varying p_{TR} . The average and standard deviation of SLAB/ESA score over 3 tests are reported for the best and last epochs. Bold face numbers represent the best performances.

	KRN-BB		KRN-SK	
	$p_{TR} = 0$	$p_{TR} = 0.5$	$p_{TR} = 0$	$p_{TR} = 0.5$
Best	0.927 ± 0.072	0.513 ± 0.102	1.117 ± 0.361	0.938 ± 0.062
Last	1.388 ± 0.494	0.943 ± 0.158	1.346 ± 0.252	1.048 ± 0.175

features that matter for regressing the bounding box corner locations are also unstable.

Interestingly, the effect of losing the local shapes and geometric features is emphasized in Table 7, which compares the SLAB/ESA scores of KRN-BB and KRN-SK for $p_{TR} = 0$ and 0.5. From the table, it is clear that when the images from PRISMA12K-TR are used during training, regressing the bounding box corners leads to consistently better performance than regressing the surface keypoints. At first glance, this is contrary to the trend in the state-of-the-art pose estimation methods based on CNNs that use surface keypoints due to their better connection to the object’s features compared to the bounding box corners. However, it is likely that such connection leads to the degrading performance as the NST pipeline inevitably disrupts the local shapes and geometric features of the spacecraft to which the surface keypoints are tightly connected. Evidently, it is difficult even with human eyes to locate all the visible surface keypoints from the images of Figure 11. Bounding box corners, on the other hand, may be instead better related to the global shape of the spacecraft which is not as damaged as the local shape by the NST pipeline.

CONCLUSION

This paper makes two contributions to the state-of-the-art in learning-based pose estimation of a noncooperative spacecraft using monocular vision. First, this work introduces a novel CNN architecture that merges the state-of-the-art detection and pose estimation pipelines with MobileNet architecture to enable fast and accurate monocular pose estimation. Specifically, the proposed CNN has mean rotation error of 3° and translation error under 25 cm when tested on the validation set, exceeding the performance of the state-of-the-art method.³² Moreover, by cropping the original image using the detected RoI and designing the CNN to regress the 2D surface keypoints, the proposed architecture shows improved performance to images of both close and large inter-spacecraft separation compared to the state-of-the-art. Second, texture randomization is introduced as a vital step in training to enable the capability of CNN architectures in space missions. The keypoint

regression network, when exposed to the texture-randomized images with 50% probability during training, results in more accurate predictions from the spaceborne images it has not seen previously. However, the analysis reveals that the neural style transfer inadvertently disrupts the local features of the target, making keypoint regression more difficult and unstable. The superior performance of the bounding box corners compared to the surface keypoints suggests that the bounding box corners are less affected by the change of local features, but their performance on the spaceborne images is still volatile nonetheless. Future work should aim at developing the texture randomization technique with minimized effect on the object’s local features.

There are few other challenges that still remain to be overcome. First of all, while the proposed CNN has real-time capability on desktop CPUs, the same on spacecraft hardware must be assessed in order to fully evaluate its applicability to an on-orbit servicing mission. Moreover, the current architecture assumes the knowledge of the spacecraft’s shape and geometry. In reality, many mission scenarios that can benefit from accurate pose estimation, especially debris removal, cannot rely on the assumption of a known target model and instead must characterize its geometry autonomously during the encounter. Future research must address the problem of robust, efficient, and autonomous characterization of the target geometry and pose determination about an unknown spacecraft or debris.

ACKNOWLEDGEMENT

The authors would like to thank the King Abdulaziz City for Science and Technology (KACST) Center of Excellence for research in Aeronautics & Astronautics (CEAA) at Stanford University for sponsoring this work.

REFERENCES

- [1] J. L. Forshaw, G. S. Aglietti, N. Navarathinam, H. Kadhemi, T. Salmon, A. Pisseloup, E. Joffe, T. Chabot, I. Retat, R. Axthelm, and e. al., “RemoveDEBRIS: An in-orbit active debris removal demonstration mission,” *Acta Astronautica*, Vol. 127, 2016, p. 448–463, 10.1016/j.actaastro.2016.06.018.
- [2] B. Sullivan, D. Barnhart, L. Hill, P. Oppenheimer, B. L. Benedict, G. V. Ommering, L. Chappell, J. Ratti, and P. Will, “DARPA Phoenix Payload Orbital Delivery System (PODs): “FedEx to GEO”,” *AIAA SPACE 2013 Conference and Exposition*, 2013, 10.2514/6.2013-5484.
- [3] B. B. Reed, R. C. Smith, B. J. Naasz, J. F. Pellegrino, and C. E. Bacon, “The Restore-L Servicing Mission,” *Aiaa Space 2016*, 2016, 10.2514/6.2016-5478.
- [4] S. D’Amico, M. Benn, and J. L. Jørgensen, “Pose estimation of an uncooperative spacecraft from actual space imagery,” *International Journal of Space Science and Engineering*, Vol. 2, No. 2, 2014, p. 171, 10.1504/ijspase.2014.060600.
- [5] S. Sharma, J. Ventura, and S. D’Amico, “Robust Model-Based Monocular Pose Initialization for Noncooperative Spacecraft Rendezvous,” *Journal of Spacecraft and Rockets*, 2018, p. 1–16, 10.2514/1.a34124.
- [6] V. Capuano, K. Kim, J. Hu, A. Harvard, and S.-J. Chung, “Monocular-Based Pose Determination of Uncooperative Known and Unknown Space Objects,” *69th International Astronautical Congress (IAC)*, 2018.
- [7] S. Sharma and S. D’Amico, “Comparative assessment of techniques for initial pose estimation using monocular vision,” *Acta Astronautica*, Vol. 123, 2016, p. 435–445, 10.1016/j.actaastro.2015.12.032.
- [8] S. Tulsiani and J. Malik, “Viewpoints and keypoints,” *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, 10.1109/cvpr.2015.7298758.
- [9] H. Su, C. R. Qi, Y. Li, and L. J. Guibas, “Render for CNN: Viewpoint Estimation in Images Using CNNs Trained with Rendered 3D Model Views,” *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, 10.1109/iccv.2015.308.
- [10] P. Poirson, P. Ammirato, C. Fu, W. Liu, J. Kosecka, and A. C. Berg, “Fast Single Shot Detection and Pose Estimation,” *CoRR*, Vol. abs/1609.05590, 2016.

- [11] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab, "SSD-6D: Making RGB-Based 3D Detection and 6D Pose Estimation Great Again," *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, 10.1109/iccv.2017.169.
- [12] M. Sundermeyer, Z.-C. Marton, M. Durner, M. Brucker, and R. Triebel, "Implicit 3D Orientation Learning for 6D Object Detection from RGB Images," *The European Conference on Computer Vision (ECCV)*, September 2018.
- [13] A. Kendall, M. Grimes, and R. Cipolla, "PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization," *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, 10.1109/iccv.2015.336.
- [14] S. Mahendran, H. Ali, and R. Vidal, "3D Pose Regression Using Convolutional Neural Networks," *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, 10.1109/iccvw.2017.254.
- [15] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes," *Robotics: Science and Systems XIV*, 2018, 10.15607/rss.2018.xiv.019.
- [16] Y. Li, G. Wang, X. Ji, Y. Xiang, and D. Fox, "DeepIM: Deep Iterative Matching for 6D Pose Estimation," *Computer Vision – ECCV 2018 Lecture Notes in Computer Science*, 2018, p. 695–711.
- [17] M. Rad and V. Lepetit, "BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects without Using Depth," *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, 10.1109/iccv.2017.413.
- [18] B. Tekin, S. N. Sinha, and P. Fua, "Real-Time Seamless Single Shot 6D Object Pose Prediction," *CVPR*, 2018.
- [19] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield, "Deep Object Pose Estimation for Semantic Robotic Grasping of Household Objects," *CoRR abs/1809.10790*, 2018.
- [20] Z. Zhao, G. Peng, H. Wang, H. Fang, C. Li, and C. Lu, "Estimating 6D Pose From Localizing Designated Surface Keypoints," *ArXiv*, Vol. abs/1812.01387, 2018.
- [21] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao, "PVNet: Pixel-wise Voting Network for 6DoF Pose Estimation," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Oral*, 2019.
- [22] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, 10.1109/cvpr.2017.690.
- [23] 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, No. 2, 2008, p. 155–166, 10.1007/s11263-008-0152-6.
- [24] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *CoRR*, Vol. abs/1804.02767, 2018.
- [25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv preprint arXiv:1512.03385*, 2015.
- [26] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, Vol. 60, No. 2, 2004, p. 91–110, 10.1023/b:visi.0000029664.99615.94.
- [27] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Readings in Computer Vision*, 1987, p. 726–740, 10.1016/b978-0-08-051581-6.50070-2.
- [28] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, "Model Based Training, Detection and Pose Estimation of Texture-Less 3D Objects in Heavily Cluttered Scenes," *Computer Vision – ACCV 2012 Lecture Notes in Computer Science*, 2013, p. 548–562.
- [29] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother, "Learning 6D Object Pose Estimation Using 3D Object Coordinates," *Computer Vision – ECCV 2014 Lecture Notes in Computer Science*, 2014, p. 536–551.
- [30] S. Sharma, C. Beierle, and S. D'Amico, "Pose estimation for non-cooperative spacecraft rendezvous using convolutional neural networks," *2018 IEEE Aerospace Conference*, March 2018, pp. 1–12, 10.1109/AERO.2018.8396425.
- [31] J.-F. Shi, S. Ulrich, and S. Ruel, *CubeSat Simulation and Detection using Monocular Camera Images and Convolutional Neural Networks*, 10.2514/6.2018-1604.
- [32] S. Sharma and S. D'Amico, "Pose Estimation for Non-Cooperative Rendezvous Using Neural Networks," *2019 AAS/AIAA Astrodynamics Specialist Conference, Ka'anapali, Maui, HI*, January 13-17 2019.
- [33] S. D'Amico, P. Bodin, M. Delpech, and R. Noteborn, "PRISMA," *Distributed Space Missions for Earth System Monitoring Space Technology Library* (M. D'Errico, ed.), Vol. 31, ch. 21, pp. 599–637, 2013, 10.1007/978-1-4614-4541-8_21.

- [34] M. A. Fischler and R. C. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” *Communications of the ACM*, Vol. 24, No. 6, 1981, p. 381–395, 10.1145/358669.358692.
- [35] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, “ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness,” *International Conference on Learning Representations*, 2019.
- [36] P. T. Jackson, A. A. Abarghouei, S. Bonner, T. P. Breckon, and B. Obara, “Style Augmentation: Data Augmentation via Style Randomization,” 2018.
- [37] M. Grant and S. Boyd, “CVX: Matlab Software for Disciplined Convex Programming, version 2.1,” <http://cvxr.com/cvx>, Mar. 2014.
- [38] M. Grant and S. Boyd, “Graph implementations for nonsmooth convex programs,” *Recent Advances in Learning and Control* (V. Blondel, S. Boyd, and H. Kimura, eds.), Lecture Notes in Control and Information Sciences, pp. 95–110, Springer-Verlag Limited, 2008. http://stanford.edu/~boyd/graph_dcp.html.
- [39] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 4510–4520, 10.1109/CVPR.2018.00474.
- [40] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” 2017. cite arxiv:1704.04861.
- [41] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, 10.1109/cvpr.2016.91.
- [42] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, “Generalized Intersection over Union,” June 2019.
- [43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” *Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1106–1114,.
- [44] T.-Y. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: Common Objects in Context,” *ECCV*, 2014.
- [45] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *CoRR*, 2014.
- [46] X. Huang and S. Belongie, “Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization,” *ICCV*, 2017.
- [47] T. Tieleman and G. Hinton, “Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude,” COURSE: Neural Networks for Machine Learning, 2012.
- [48] I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization,” *International Conference on Learning Representations*, 2019.
- [49] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, “Random Erasing Data Augmentation,” *arXiv preprint arXiv:1708.04896*, 2017.