EuroGNC Navigation Guidance Control CEAS Bristol, UK June 11th-13th 2024
University of BRISTOL ROYAL AERONAUTICAL SOCIETY AIAA

A Review of State-of-the-Art 6D Pose Estimation and its applications in Space Operations

Siddharth Singh PhD Researcher, Cranfield University, Aerospace, MK43 0AL, Bedford, UK. Siddharth.Singh.026@cranfield.ac.uk

Hyo Sang Shin Professor of Guidance, Navigation and Control, Cranfield University, SATM, MK43 0AL, Bedford, UK / Professor, Cho Chun Shik Graduate School of Mobility, KAIST, Daejeon 34141, Republic of Korea. h.shin@cranfield.ac.uk/

Antonios Tsourdos Director of Research for SATM / Head of the Centre for Autonomous and Cyber-Physical Systems, Cranfield University, SATM, MK43 0AL, Bedford, UK. a.tsourdos@cranfield.ac.uk

Leonard Felicetti Lecturer in Space Engineering, Cranfield University, SATM, MK43 0AL, Bedford, UK.leonard.felicetti@cranfield.ac.uk

ABSTRACT

Increase in autonomous systems now requires for these systems to work in close proximity of other objects in their environments, with many tasks that need to be done on environment objects for eg., assembly, transportation, rendezvous, docking, or to avoid them like collision detections/avoidance, path planning etc. In this literature review we discuss machine learning based algorithms that solve the first step of vision-based autonomous systems i.e., vision based pose estimation. This paper presents a critical review in advancements of 6D pose estimation using both 2D and 3D input data, and compare how they deal with the challenges shared by the computer-vision based localisation problem. We also look over algorithms with their applications in space based tasks like in-orbit docking, rendezvous and the challenges that come with space-vision applications. To conclude the review we also highlight niche problems and possible avenues for future research.

Nomenclature

DL	=	deep learning
ROI	=	Region Of Interest
CV	=	computer vision
CNN	=	convolutional neural network
GNC	=	Guidance Navigation and Control
GAN	=	Generative Adversarial Networks
ICP	=	Iterative Closest Point
PnP	=	Perspective-n-point
PE	=	pose estimation
$RANSAC$	=	random sample consensus
$FLOP$	=	Floating Point Operations

NN = Neural Network
$SE(3)$ = Special Euclidean group of 3D rigid body displacements

1 Introduction

6 Degree of Freedom (6DOF or 6D) pose estimation (PE) is the first step to many autonomous vision-based systems and an essential field of research in Computer Vision (CV) applications like robot manipulation [1], autonomous driving [2], construction [3] etc. The 6D information that we try to extract from a target object are its coordinates in *x, y, z* and its orientation in terms of *pitch, roll and yaw* with respect to the camera coordinate system. Machine learning solutions can use different kinds of inputs to solve the pose estimation problem with 2D or 3D information i.e., RGB, RGBD and point cloud inputs. These different inputs lead way for diverse applications and methodologies but share problems that are associated with vision systems. All input information, 2D or 3D, face challenges of auto-occlusions, inter/intra-category symmetry, lighting, unusual views, noise in input and generalisation over unseen objects[4–6].

Pose Estimation has been a research section in computer vision for a long time, with older solutions being geometrical inference based, i.e., they try to establish 3D-2D correspondences to regress 6D pose of object. The best example for this is *Key-point, ICP [7], RANSAC [8]*, which use extracted points or features to solve for pose using feature mapping and matching. Simple in logic and implementation these methodologies suffer from low convergence speeds and accuracy, as well as difficulties and high computation times with complex geometrical features, thus these algorithms require a close initial estimate making them impractical for real-time application but are still used as a last step refinement module.

With the recent growth of Deep Learning (DL) CV techniques, prior methods were made obsolete by introduction of Template-based methods, which assess the target object as multiple representation from different views and compare them against the observed object. This approach does provide high accuracy in inter-class pose estimation but falls short to the occlusion, lighting and unusual view problems faced by vision based approaches. To achieve robustness in such methodologies the object representation data-set needs to be increased in size, diversity and the number of comparison, to give high accuracy results but these increments lead to increase in computation demand and thus inference speed. Understanding these short-comings DL in computer vision also introduced several other methodologies like Learning-based methods, Bounding Box / Key-point and Perspective-n-Point (PnP) based, Regression based, Latent Learning etc. The building block for these techniques is the introduction of feature mapping between extracted features and 6D pose information using Convolutional Neural Networks (CNNs). The PnP[1] algorithm has been used extensively in feature mapping between 2D features and 3D models of objects and it does so by solving the camera projection problem. In Bounding box / Key-point based methods CNNs predict a 2D to 3D bounding box and these bounding boxes are compared with the 3D CAD model using the PnP algorithm. Other types of Learning based methods like Regression, which use a neural network made up of multiple CNNs to directly regress the pose of an object, are very fast in inference but require a large and diverse data-set to train upon and perform poorly on domain gaps as well extreme or unseen conditions.

6D PE is also a challenge in space applications and operations as most control systems in space technology are mostly remote-controlled and are moving towards autonomous systems given the harsh environmental challenges of space and the increase in communication / response time over astronomical distances. The necessity 6D PE comes because of close proximity operations like docking, rendezvous, de-orbiting satellites externally, in-orbit assembly and collision avoidance. Most space applications use LIDARs [9] for depth mapping (3D) as well as RGB input to provide high accuracy of pose and depth estimation. But, new solutions are aimed to provide same high accuracy and fidelity information using 2D images in an approach to get rid of LIDARs and the technical points of failure that they come with as well as the

[1] https://jingnanshi.com/blog/pnp_minimal.html

respective weight and cost which are primary consideration in any space operation.
In this review paper we cover multiple approaches for the 6D pose estimation challenge using advancements in machine learning and Deep Learning, with a highlight on similar implementations and solutions that are in use for space applications. We cover major advancements in Pose Estimation with all kinds of inputs i.e., RGB, RGB-D and Point Clouds. In conclusion we highlight the niche problems faced by Vision-based systems still and propose new avenues for research and solution methodologies for 6D Pose Estimation.
The paper is structured as follows : Section 2 explains the selection-ideology behind papers chosen for review, Section 3 illustrates 2D image based solutions for Pose Estimation with subsections 3.1 covering Render based methods, 3.2 highlighting Key-point/Bounding Box methods and 3.3 focusing on Learning-Based methods. Section 4 discusses methodologies with RGB-D and Point Cloud inputs. Section 5 introduces the applications of similar solutions for space-based operations. Section 6 concludes the paper with a highlight on the niche problems and short comings of the papers reviewed and possible directions of future research.

2 Selection Methodology

To keep the survey concise while considering all forms of input data, we set rules to identify potential papers for review and reject any outliers. The criterion's set for paper selection have been described in Fig.1, showcasing the PRISMA literature review collection.

1) To keep the research up-to date, papers published in 2016 and after will be considered as this was the boom in the use of DL in CV and PE challenges.
2) We select papers that introduce new techniques or metrics rather papers that work on optimising an already present solution. Example, EPro-PnP [10] which improves upon its previous v1 architecture by adjusting network weights and initialisation, in the improved v2 publication.
3) Papers to be considered need to be implementable in real-time operations, Pose estimation in GNC needs to have fast real-time inference speeds.
4) In our selection of Point Cloud based methods we discuss pose estimation techniques as well as state-of-the-art point cloud completion tasks to better understand the integration of 3D information and geometry based operations.
5) Keywords used for search - '6D Pose Estimation', '3D Localisation and Tracking', 'Pose Estimation', 'Spacecraft Pose Estimation', 'Point Cloud Registration', 'Point Cloud Pose Estimation', 'Vision based robotic control'- were used for searching and refining papers from different publications.
6) Vision based research papers have been taken from the IEEE Computer Vision and Pattern Recognition Conference (CVPR)[2], ArXiv[3], ScienceDirect[4], MPDI[5]

2.1 Improvement in Computational Resources

We also take into account the change in computational strength of processors and new possible application with GPU accelerated hardware. Lentaris Et al.[11] goes over a critical review of embedded processing systems for space applications. For an embedded system to be certified space-worthy there are a few import criteria such as radiation resistance, working temperature envelop amongst many so that the system can operate fully in harsh conditions of space. In [11], we can see the highest performance of an embedded system is capped in the GFLOPS (Giga/Billion Floating Point Operations) range. The highest performance metric we see in terms of FLOPS is 900 MFLOPS(Million FLOPS) in CPU systems

[2]https://cvpr.thecvf.com/
[3]https://arxiv.org/
[4]https://www.sciencedirect.com
[5]https://www.mpdi.com/

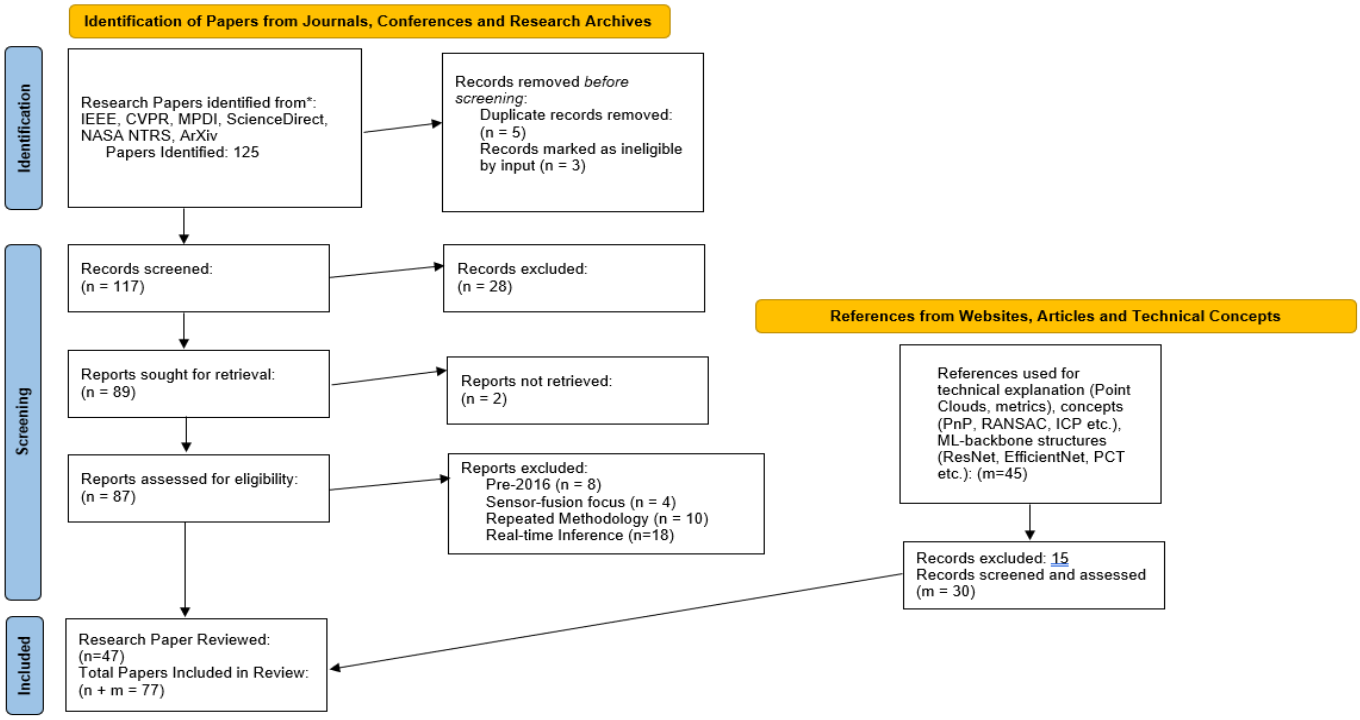


Fig. 1 PRISMA systematic Review

and a high of 38 GFLOPS in FPGA based system architecture. AiTech[6] have verified the NVIDIA Jetson TX2i[7] is space worthy with radiation resistance, suitable temperature and G-force envelop. This is a giant leap in computational strength possible for space applications as the TX2i-Nano being a smaller computing unit still provides 1.3 TFLOP (Tera/Trillion FLOPS) which is a minimum of 100X better than previous hardware[8]. With such computational strength it is possible to implement more accurate and flight worthy AI systems on board for autonomous operations. Heavy machine learning algorithms usually require computing strength in the 10-100 GFLOPS range for real-time inference, which leaves us vast amounts of computational strength for stronger and better navigation and control systems further down the line.

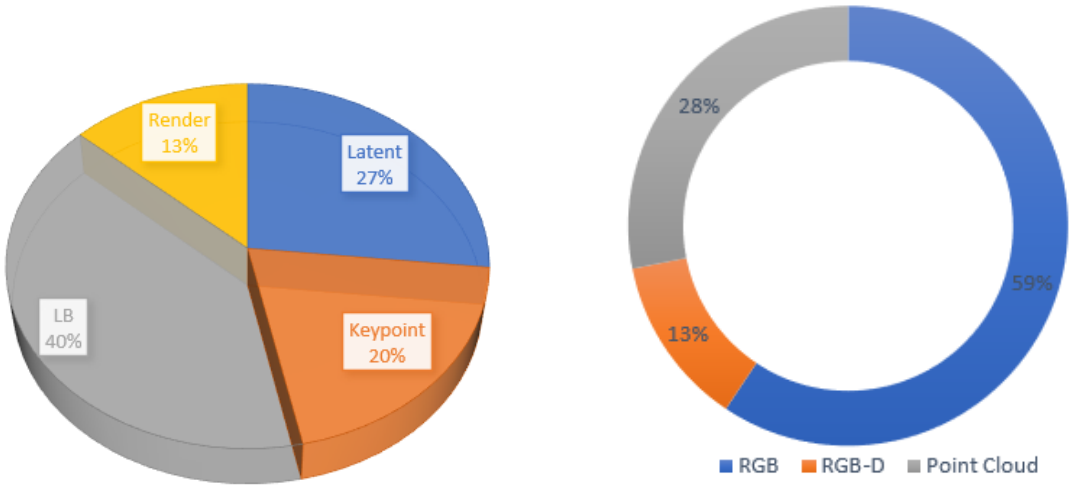
The conclusion of the selection leaves us with a total of 47 research papers and their distribution according to methodology and input is show in Fig.2, and Fig.3 shows the overall distribution of papers and their subcategories.

3 Pose Estimation using 2D Input : RGB

3.1 Template/Render-based Methods

These approaches also called template-based methods due to the creation of a database using either active or passive rendering [6] and are usually a multi-stage process with multiple neural networks in series and an active rendering system that runs internally with the neural network architecture or is saved as a database. The approach can be broken down in to a few simple steps. The first step is extracting image features from the input using a feature extractor like ResNet [12], to extract 2D correspondences or features and semantic maps. These maps are compared with images of a 3D CAD model rendered

[6]https://aitechsystems.com/space-products/
[7]https://developer.nvidia.com/embedded/jetson-tx2i
[8]https://www.eenewseurope.com/en/jetson-gpu-space-qualified-for-ai/



(a) Methodology based - Latent, Render, Key-point, LB(Learning Based)

(b) Input Based - RGB, RGB-D, Point Cloud

Fig. 2 Distribution of collected papers based upon methodology *(left)* and Input *(right)*

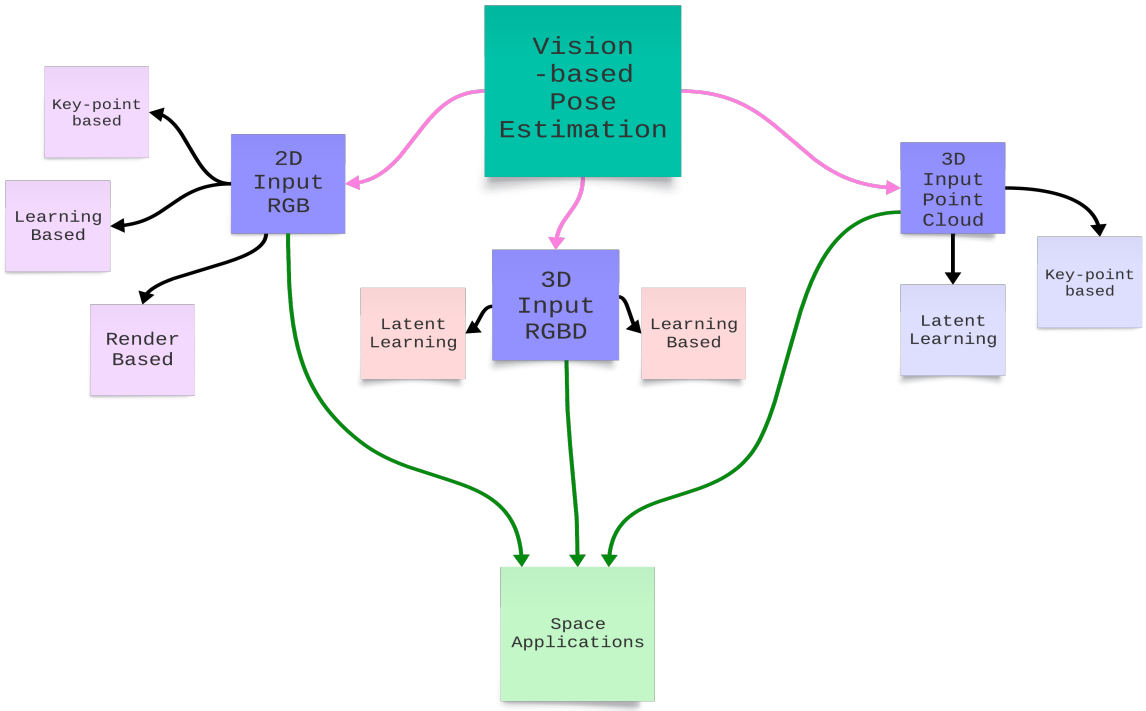


Fig. 3 Mind map of review distribution

online, to find the closest match and predict pose as SE(3) using the render properties and similarity. DPOD [12] uses PnP and RANSAC to compute the initial pose between these correspondences and is further fed into a RGB-based pose refinement module. The pose refinement module compares rendered pixel values and observed pixel values to further refine the initial pose estimate. DEEPIM [13] is a pose refinement module with its prerequisite being an available 3D CAD model and an initial pose estimate. It uses the input image, rendered image and mask and an initial pose to form a feature map which is further fed to multiple CNN's to predict pose and translation separately.

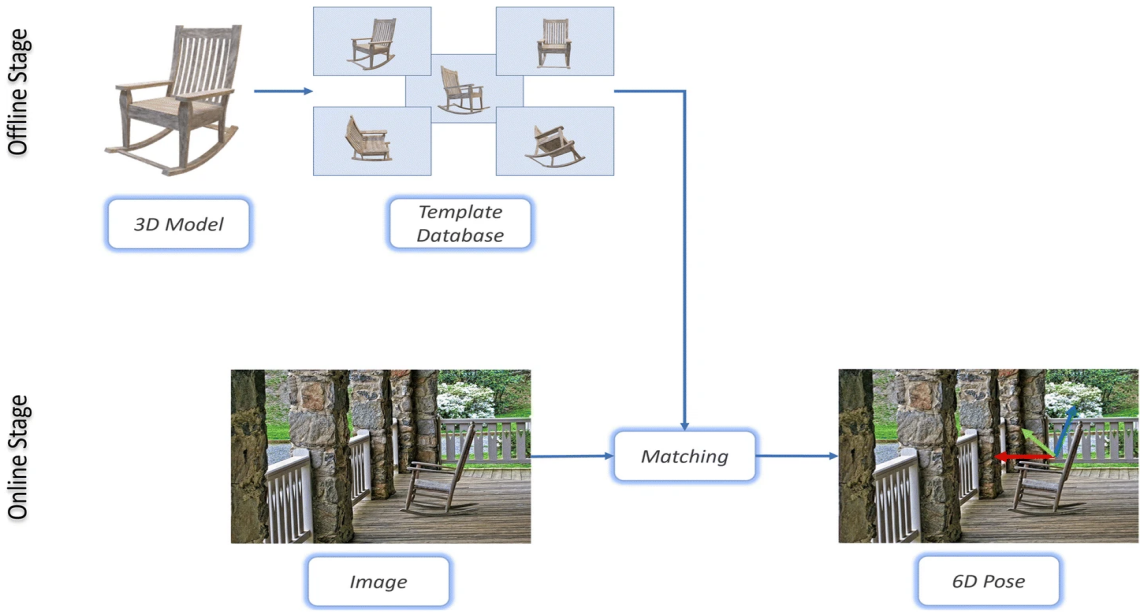


Fig. 4 Schematic representation of Render-based approaches [6]

These methodologies come with their own pros and cons, as they perform well with exhaustive 3D CAD based data sets and for texture-less objects but are highly sensitive to lighting or occlusions as This type of disturbance or noise is difficult to be included in Render-based models, and they change the similarity index between rendered image for pose estimation versus the observed image. The repetitive task of rendering images for each cycle of inference makes these algorithms computationally slow and not easily applicable in real-time environments. To improve upon this computation cost of rendering instances ImplicitAAE [14] introduced use of Augmented Auto-encoders that use active rendering to train the auto-encoder to reconstruct SE(3) pose of the object from randomly generated images of the CAD model, and the encoder outputs are saved into a code-book of representations and respective orientations. The trained encoder-decoder architecture is given the observed image and semantic map to create a representation and use cosine similarity to regress the pose of the object with an optional depth map based Iterative Closest Point(ICP)-pose refinement module, to increase accuracy. Though it does not use active-rendering it uses a database based on embeddings generated by the encoder thus also referred to as Template-based methods. Marulo and Tanzi [6] go into further detail of differentiation for template based methodologies, and are not considered for this review due to their higher computation times and lack of implementation in real-time applications. RNNPose [15] improves upon previous render based methods by including 3D information by extracting features from rendered image of CAD model, initial pose orientation of the CAD model (based on a initial pose estimate) and the observed image. The features from the observed image and rendered image are combined to compare against the 3D correspondences using a Gated-recurrent-NN and differentiable LM (Levenberg–Marquardt) optimisation with rigid body constraints to update initial pose estimation with each iterative step in relative SE(3) transformations. This allows [15] to perform better in occluded and unusual view or lighting circumstances over previous methods [14], [12], [13], by keeping account of previous poses and their correspondences which helps improve accuracy during occlusion and noisy scenarios. Even with the advances, the high computation cost of rendering, variety of lighting conditions and occlusions still stand as challenges to render based approaches. There are other methodologies like CPS++ [16] and CloudAAE [14] that use a differentiable render for training but not for inference and evaluation, and will be discussed in Sec.3.3. Approaches [15], [14], [12], [13] are compared over common data-sets that have been used for training and evaluation, i.e., we compare their performance metrics over the LINE-MOD [17] and Line-MOD-Occluded [18] data-sets in Table 1 over the ADD-S metric which is an ambiguity-invariant pose error metric which takes care of both symmetric and non-symmetric objects into an overall evaluation. We will be using the

ADD-S metric to do performance comparisons in the review.

METHODS	DPOD [12]	DEEP IM [13]	Implicit AAE [14]	RNNpose [15]
YCB-V [5]	76.3	81.9	-	**83.1**
Linemod [17]	**95.15**	88.61	71.8	93.7
Occlusion [18]	47.3	-	-	**60.65**

Table 1 Comparison of Template/Render Based Methodologies over the ADD(S) metric

As we can see RNNPose [15] performs better as an overall among different datasets, with methods like [13] and [14] that cannot perform well under occlusions as they have no first hand understanding of the occlusion or the target object. Random Occlusions are the biggest challenge for Render based methodologies as any kind of noise or occlusion inclusion in the render degrades overall performance, while improving Occlusion performance slightly. DPOD [12] is also able to perform in Occluded environment using its correspondence generator before rendering the CAD model to help with pose regression. This is the overall drawback of Render-based methods, the need for clean and occlusion free-environments supplemented with pre-existing CAD models for rendering. These donot allow for robust, computationally light or accurate prediction models as the observed image maynot be a 1:1 replica of the CAD model. To improve upon this Feature-based methods try and extract image/target features to regress pose.

3.2 Feature-PnP-based Methods

Key-point/bounding box based methods depend on feature extraction from the observed image, these features can be key-points or bounding boxes. The extracted features are then compared with an annotated 3D model of the target object to regress pose using a PnP solver for 2D-3D matching. Feature based methods are usually a multi-stage pipeline with Object-detection, key-point extraction, 2D-3D matching and a pose refinement module for better accuracy. These methodologies deal with occlusions well as CNN's are exploited to predict occluded key-points and still continue with the 2D-3D mapping. A pose-refinement module is added to aid with occluded pose prediction stability [19].
These methodologies work well with occlusions but fall short when the target object is symmetric or texture-less, as the accuracy of key-points detected and the key-points themselves being unique allow for a close 2D-3D match to predict 6D coarse information. Lack of distinct features leads to lower accuracy of key-point detection and thus an overall performance degradation in pose estimation. To extract a final pose estimation the 6D coarse prediction is fed into a pose refinement module.6D Pose w/o PnP [20] takes on the challenge to do bounding box based pose estimation without using a PnP solver. To do this they define a pose and translation parameters in terms of extracted features to project the predicted 3D bounding boxes in 2D. They provide state-of-the-art results and a high computation speed of 17ms [20] and is considered a Learning based methods more-so than a Feature based method.

HybridPose [21] is a perfect example as it extracts 3 different kinds of features, Key-points, edge vectors and similarity correspondences. These features are used to form an initial prediction which is optimised using a generalised German-Mcclure (or GM) robust function [21] to solve for non-linear pose refinement.YOLOPose [22] and POET [23] are similar applications of transformer architecture for pose estimation. Both methodologies use a ResNet [24] neural network as a feature extraction backbone, after that differ in the way the outputs of the extractor are used. YoloPose [22] uses a simple architecture of encoder-decoder-prediction head using the bounding boxes, class probabilities, output embeddings or Canonical 3D points.All decoder embeddings are a set of *class probability, Rotation estimate, translation*

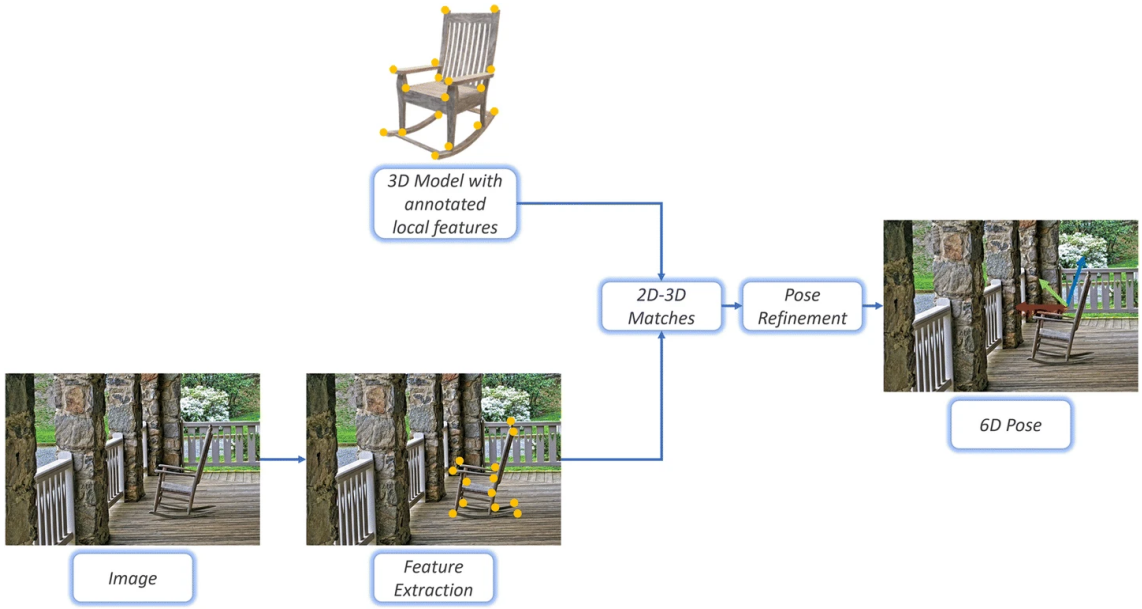


Fig. 5 Schematic representation of Feature-based approaches [6]

estimate, key-points and bounding boxes, which are fed into a FFN to predict pose and optimise network by using different losses for all outputs of the set, with a conditional pose loss for symmetrical object. [23] on the other hand uses extracted bounding box to incorporate positional encoding for dynamic object queries to the decoder, while using the features extracted for attention generation. The output of the decoder is fed into another FFN to separate translation and rotation for each object query. PrimA6D [19] is another fully transformer based architecture that uses a Variational Auto-Encoder (VAE) based encoder decoder structure with 2 separate decoders for primitive(axis) reconstruction and an object reconstruction. Both the decoders are trained using a GAN methodology. Key-points are extracted from the reconstructed primitive using a [24] feature extractor and pose is regressed using a PnP solver on the key-points extracted. It improves upon previous methods by ignoring the complexity of the 3D structure and converting it down to a primitive-axis-representation which allows for more accurate and fast PnP or ICP solution. This primitive-axis-representation deals with occlusions and disturbances robustly as its more accurate to predict 3 points in a mathematical relationship than occluded key-points. [25] utilises a YOLO [26] architecture for key-point detection to estimate pose using a given 3D CAD model of the target object using PnP. In their study they summarise with great performance in 2D localisation but fall short in depth and yaw estimation as a short coming of 2D image information, verifying these results and drawbacks with flight experimentation. [19–23] are compared in terms of performance metrics over the common data-set LINEMOD [17, 18] for [19–21] and the YCB-V [5] data-set for [22, 23] in Table 2.

METHODS	Prima6D [19]	YOLOPose [22]	POET [23]	HybridPose [21]	w/o PnP [20]
YCB-V [5]	**94.43**	91.2	92.8	-	-
Linemod [17]	**97.62**	-	-	91.3	92.68
Occlusion [18]	**79.2**	-	-	74.5	-

Table 2 Comparison of Feature (Key-point/Bounding Box) Based Methodologies over the ADD(S) metric

As we can see in Table 2, [19] out-performs all other methodologies over every dataset. This is due to its ability to deal with complex geometries and occlusions in a simple yet robust manner using primitive representations. Transformer based models deal better with symmetrical objects with the help of positional encoding, and attention based geometry learning. With the boom in Deep Learning and

speed of computation solutions like [22, 23] try to bridge the gap between Feature based methods and purely Learning based methods with the removal or reducing the computation load for inference based pose refinement steps like PnP, RANSAC or ICP.

3.3 Learning Based Methods

Learning based methods try to regress the pose of target object directly from the input image. Since these methods depend completely on CNN's they require high amounts of training data and training time to provide good accuracy with faster inference times. Learning based methods can have either one-stage pipelines like [27–29] or multi-stage pipelines like [5, 10, 30]. While One stage algorithms try to predict pose directly by learning image features and their co-relation to output by using heat-maps and correspondences, multi-stage networks incorporate different extracted features like key-points, colour-coordinate maps, bounding boxes etc. to regress a final pose.

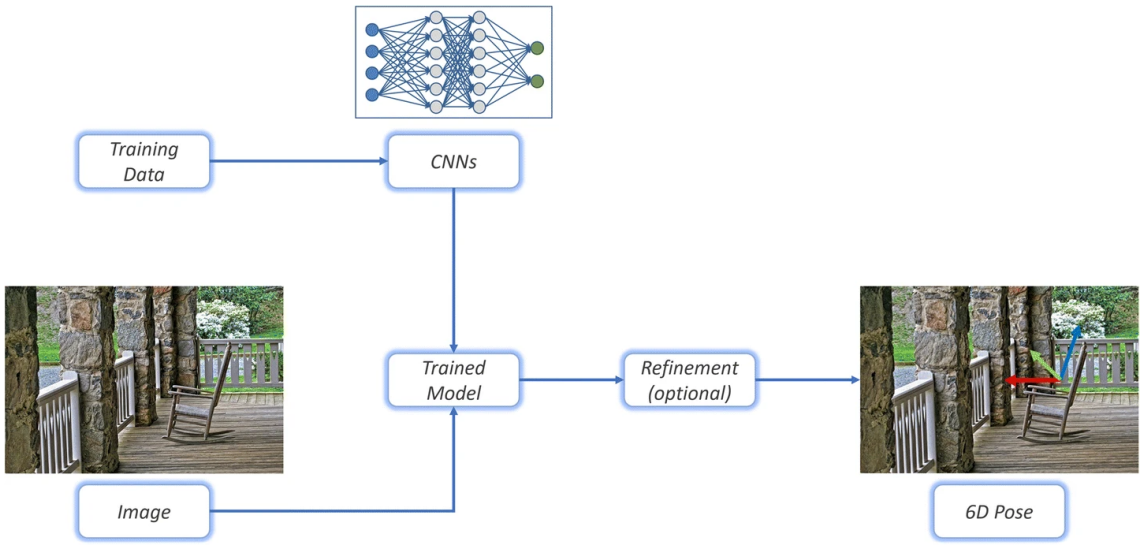


Fig. 6 High level representation of Learning based methods [6]

PoseCNN[5] handles the task of 6D pose estimation by combining outputs of three different tasks purely using CNN's. It carries out semantic labelling, translation regression and rotation regression and then incorporates these outputs with respective Region of Interest (ROI) to regress the final pose. It is robust to occlusions and lighting by combining the outputs of semantic, translation and ROI predictions to predict the final pose, giving state of the art results on the [5] and [18] data-sets. The performance of the PoseCNN algorithm increases with an additional ICP based pose refinement module. YOLO6D [28] takes the challenge one step further by proposing a CNN architecture for single-shot 6D pose estimation. The pipeline extracts object centroids, 8 bounding box points, and using these 9 extracted or predicted points as control points, [28] solves a PnP algorithm for pose estimation all in a single shot without any pose-refinement module giving it a high inference speed of 20ms. Similarly as discussed in 3.2, [20] uses a similar CNN structure with the novel introduction of a Collinear equation layer and a shared backward propagation process to do avoid using PnP and thus giving it its high speed inference. Pix2Pose [27] improves the methodology for 2D-3D correspondences by converting a 3D CAD model to a normalised-coordinate colour coded image, and uses a one-stage CNN pipeline to predict the the same and a error prediction from the input image. It uses the predicted normalised map and predicted error map with a iterative PnP-RANSAC solver to predict pose. It also leverages upon the advancements in GAN's for training using synthetic data. A drawback being the need for bounding box detector module to provide the model with necessary inputs. SSD6D and BB8 [30, 31] on the other hand gets rid of the multiple re-sampling of the image by extending the Visual Geometry Group (VGG) backbone [32]

to directly regress class, pose and respective confidence scores. Similarly EfficientPose[29] build upon extending the EfficientNet-BiFPN (Bidirectional Feature Pyramid Network) [33] with addition of rotation and translation prediction heads (networks). It also carries out an iterative pose refinement using depth-wise-separable convolutional layers, thus avoiding the computationally expensive PnP, ICP or RANSAC refinement techniques.
With learning based methods, a lot of annotated training data is required and still has low generalisation and the PnP solution has been a challenging one for complex, large structures. To improve upon these [10] proposes a probabilistic PnP layer for general end-to-end pose estimation, which outputs a distribution of pose with differentiable probability density on the SE(3) manifold. The 2D-3D coordinates and corresponding weights are treated as intermediate variables learned by minimising the KL divergence [34] between the predicted and target pose distribution.

METHODS	EfficientPose [29]	PoseCNN [5]	Pix2Pose [27]	BB8 [30]	SSD6D [31]
YCB-V [5]	-	75.9	-	-	-
Linemod [17]	**97.35**	-	72.4	89.3	90.37
Occlusion [18]	**83.98**	78	32	43.6	55.95

Table 3 Comparison of Learning Based Methodologies over the ADD(S) metric

Compared to [5, 27, 30, 31] that focus on a close initial pose-approximate, to reduce the computation time and load required by modules like RANSAC, PnP and ICP. EfficientPose [29] utilises only image features and iterative NN-based refinement to predict and refine initial pose estimate thus being computationally lighter and a more accurate pose-refinement. The EfficeintDet [33] backbone utilises feature propagation and ROI based multiple predictions, allowing it to deal with occlusions well using local and global image features. The trade-off lies between the number of pose-refinement steps and accuracy with inference speed. The EfficeintPose model proves robust by performing well with a single refinement iteration as shown in metrics of 3.

4 Pose Estimation using 3D Input

4.1 RGB-D Input

As the name suggests RGB-D methods include a depth mask or map as an input to the system. It greatly improves performance as depth is an essential information that monocular vision systems cannot extract accurately. With the addition of depth as an addition channel of information, RGB-D methods focus on information fusion between depth and RGB treating them as individual and independent sources of information. DenseFusion [35] introduces this methodology with separate feature extraction networks for depth and RGB information, to extract depth based point cloud embedding and RGB based colour embeddings (instance segmentation), and then fuses them in a pixel-wise manner with a global average feature, for per-pixel prediction of *rotation, translation and confidence* and giving out final results using non-max suppression[9] and an iterative refinement module. The research highlights the significant improvement in performance *(over 10%)* over SOTA RGB based methodologies[35–37]. In the same year,2019, MaskedFusion [36] proposed a similar methodology about feature fusion between RGB and Depth information, it improves upon [35] by improving the feature extraction and fusion step by extracting 3 different feature sets from depth, RGB and segmentation mask, to create a complete feature set while ignoring unnecessary information. This feature set is used to predict an initial pose and use a refinement module to give the final output. It improves its performance significantly over [35] by the application of

[9]https://learnopencv.com/non-maximum-suppression-theory-and-implementation-in-pytorch/

letting go of unnecessary information during feature extraction and pose estimation. To bridge the gap between RGB-D methods and 2D methods LatentFusion [38] utilises a embedding-input active rendering system, where the network is trained to predict depth maps from latent representations of RGB and segmentation mask. The system further uses this depth prediction with RGB and mask to predict pose in an iterative way with the help of refining using the rendering system. The down-side of the algorithm being it requires multiple view points of the object to create a latent representation. This methodology acts as template-based method with an active rendering for refinement as explained in Sec.3.1. FFB6D [37] improves upon Dense-Fusion [35] by implementing bi-directional fusion modules (see Fig. 7) between the two networks (RGB and Depth/point cloud feature extractor) allowing for sharing local and global information during the encoding and decoding of features to allow for better appearance and geometry learning. The research stands as the SOTA in pose estimation against other RGB-D and RGB methods, without the need for a pose-refinement module. The fusion of information during feature extraction is done by information fusion between 2D to 3D and vice-a-versa using nearest-neighbour mapping between the 2 feature sets, while sharing prior information as well using skip-residual connections in the architecture. These features are used to predict 3D key-points of the object and predicts 6D pose as a Least-Squares Fitting problem [39]. A slightly improved and SOTA implementation can be seen in PVN3D [40].

METHODS	FFB6D [37]	LatnetFusion [38]	MaskedFusion [36]	DenseFusion [35]
YCB-V [5]	93.1	-	**93.3**	93.1
Linemod [17]	96.6	87	**97.8**	94.3
Occlusion [18]	**66.2**	-	-	-

Table 4 Comparison of RGBD Methodologies over the ADD(S) metric (Average Distance between point clouds)

RGB-D methods out-perform RGB methods purely because of the inclusion of depth information, which RGB methods have to predict or regress to perform 6D pose estimation. Table 4 compares [35–38] and highlights their increase in performance over RGB methods over the same data-sets. [37], [36], [35] all perfomr similar on the YCB-V [5] dataset, with FFB6D [37] and MaskedFusion [36] performing better on the LineMod [17] dataset. MaskedFusion reaches this high accuracy performance by combining features of RGB and Point Cloud using the [41] architecture to introduce structural information during PE, with FFB6D following the [29, 33] methodology by feature-fusion as shown in Fig.7 *bottom*. This also allows for good performance under-occlusion but still not at par with earlier discussed 2D methodologies, as occlusions remove part of the 3D information, and these algorithms do-not include complete geometrical structural learning, as they treat RGB-D as two separate channels of information rather than a single data structure. Fig. 7 highlights the overall ideology behind RGB-D methods using the Dense-Fusion [35] and FFB6D [37] network architectures.

4.2 Point Clouds

Point Cloud based approaches differ from RGB-D methods as they treat point clouds i.e., x,y,z as the input information rather than separating the 2D projection and Depth information as in [35] and other RGB-D methodologies. We discuss different point cloud based tasks aside from Pose Estimation to further understand Point cloud implementations and how this information is exploited. [41–45] discuss advancements in Point cloud based tasks and transformer based latent learning methods, and [46–50] discuss SOTA pose estimation methods using 3D point cloud data.

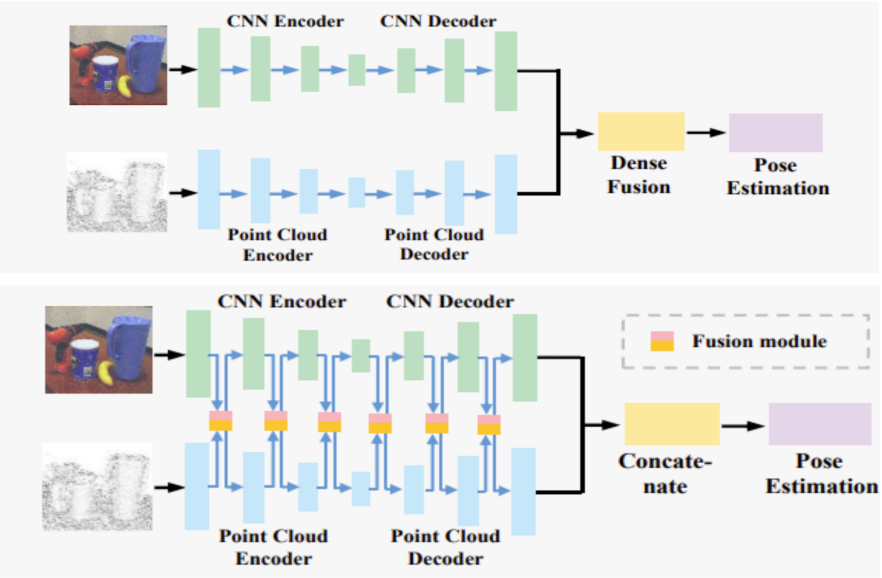


Fig. 7 ***(Top)*** **Schematic of Dense-Fusion[35] architecture representing the basic ideology behind RGB-D methods,** ***(Bottom)*** **FFB6D [37] architecture, highlighting the novel bi-directional information fusion methodology.**

4.2.1 Latent-learning with Point Clouds

Transformers in Natural Language Processing (NLP) [51, 52] introduced the concept of positional encoding and context understanding with local and global features over the complete sentence. This same idea was implemented with Point Clouds to treat them as an ordered set of data. PointNet [41] and the use of T-Nets which are special NN's based on Spatial Transformer Networks, which learn consistent transformations between point clouds, introduce converting point-clouds as features for segmentation and classification tasks. To improve upon [41], Point Cloud Transformer [42] introduces transformer architecture for point cloud representations, with attention networks allowing for relevant global and local feature collection. PCT [42] set the benchmark for transformer based point cloud methods and researches [41, 43, 44, 49] etc. PointBERT [43] and PoinTr [44] extend upon the idea of representation-learning point cloud tasks by introducing *Geometry-aware Transformers* [44] and *Masked Learning* [43]. These ideas allow avenues for representation based learning with point clouds instead of inference-heavy methods like ICP [7], RANSAC [8] and PnP [53]. As we can see in Fig.8 the PoinTr architecture focuses on rebuilding observed point cloud features using an Geometry-Aware encoder-decoder architecture, to predict the missing points in a reconstruction task Fig.8 shows the conversion of point clouds to embeddings and how they are used for latent learning using the [44] network architecture. ConDor [45] takes advantage of this latent learning to do canonicalisation of point clouds, i.e., completing point clouds and understanding the overall reference structure of the source point clouds. These studies indicate the useful application of latent/representation learning methods for point clouds and we see how they have been implemented for the pose estimation task.

4.2.2 Pose estimation with Point Clouds

PointVoteNet [49] utilises the PointNet [41] architecture for part-segmentation to predict pose via a scoring and voting system using a CAD model as reference.PointPoseNet [50] introduces a transformer architecture to predict 3D point wise vectors using extracted bounding boxes from a [26] backbone, utilising point-wise 3D vector prediction based on the part-segmentation thus infers more local information compared to [49]. It allows the PointPoseNet to predict hidden and invisible 3D key-points and optimise selection by using a preemptive RANSAC and *Furthest Point Sample*. Cloud AAE [47] implements the

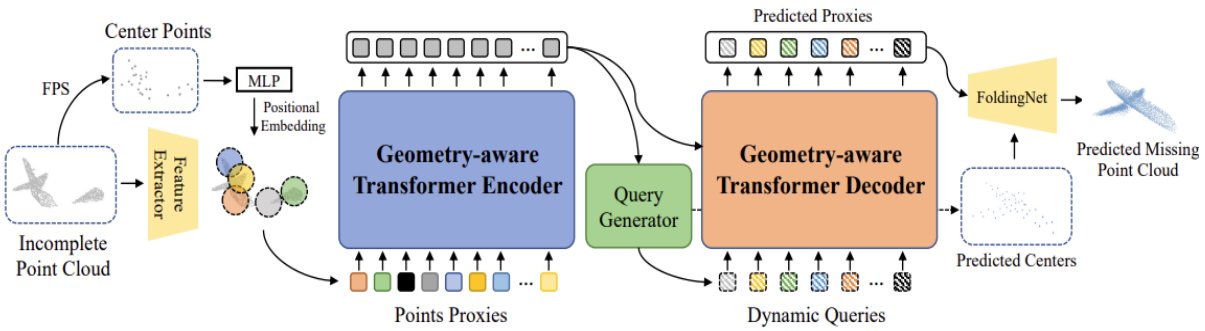


Fig. 8 The PoinTr[44] architecture representing the uses of latent learning for point cloud completion tasks, i.e., generating and completing incomplete embeddings to recreate the point cloud

first latent information based pose and translation prediction system by using Augmented Auto-Encoders (AAE) to create latent embeddings of a point cloud, and use Multilayer Perceptron (MLP) for pose and translation prediction. Plum [48] is an introduction of reward based look-up method for point cloud registration, and uses a reward based optimisation rather than a match-loss metric making it more robust than other methods like [7] and its variants. It stands as a bridge between inference based pose estimation methods and template-based deep learning methods, providing state of the art results and computational speeds. To improve upon the idea of latent learning, Query6DOF [46] implements latent embedding comparison and regression to predict pose using point clouds. It does so by creating a code-book of point cloud embeddings and their respective pose information, and compare incoming observed embeddings to find the nearest match and refine the respective pose. It is an implementation of Template/database type methodology but reduces computation load, thus increasing inference speed as its is comparing latent embeddings, rather than complete point clouds. A drawback of the system is the need for immense training data to create a sufficient data-base that can perform in robust conditions of occlusion, unusual views and complex 3D structures.

METHODS	PVN3D [40]	CloudAAE [47]	PointPoseNet [50]	PointVoteNet [49]
YCB-V [5]	92.3	**94**	93.2	-
Linemod [17]	**99.4**	95.5	98.4	96.3
Occlusion [18]	-	66.1	**79.5**	52.6

Table 5 Comparison of Point Cloud based Methodologies over the ADD(S) metric (Average Distance between point clouds)

Table 5 shows how [47] is the simplest implementation of latent-learning based PE, an encoder -decoder architecture than regresses pose directly from the latent embeddings, providing comparative results to SOTA RGB methods in Sec.3, specially under occlusion. More advanced methods like PointPoseNet [50], perform better overall and significantly better under occlusion because of its point cloud based key-point prediction based on unit-vector predictions and pose matching using least-square fitting. The comparatively low performance in occlusion to 2D based methods is because of the lack of complete 3D information. To improve performance combination of point cloud reconstruction and then pose estimation may provide better results for occluded and clear views. Point cloud based pose-estimation methods are not widely used because of their computational expense, but with deep-latent learning based methods these computational needs can be reduced and used for real-time applications.

5 Pose Estimation in Space Applications

Most pose estimation methodologies in space applications like to depend on 3D information using LIDARS as 3D information based estimation is superior to 2D information. Most space based systems usually have a LIDAR or some form of laser range scanner[54–56]. There are 2D methodologies that have been proposed for uncooperative pose estimation. [55] major contribution was on the use of Photonic-Mixer Device (PMD) Camera that allows for 3D scan matching with the model of the space craft available. The paper does pose estimation as a Least Square Fitting Problem [39], focusing on the calibration and data-processing from the PMD camera. Cassinis and Fonod [57] highlights the use model-based matching and use of inference level algorithms like ICP, PnP and RANSAC with either direct 3D input or synthesised features from 2D images. Pesce and Opromolla [58] is an example as it extracts major features from a 2D image and uses RANSAC aided with PCA (Principal Component Analysis) and solves for pose using a EPnP solver [59]. There are multiple 2D methodologies that try to regress pose directly from 2D image feature maps like, [60–62] are methods that regress pose directly from feature maps extracted from a 2D image using pre-trained networks[24, 63] for their high accuracy and performance over generalised vision based tasks.

Researchers also try to overcome the the lack of annotated data in researches [64–66] that introduce synthetic data generation platforms using open source software like Blender and Unreal Engine. URSONet[65] focuses on creating a realistic data-generation platform that overcomes the domain gap between synthetic and real data. They use a CNN backbone for feature extraction and regress translation in x,y,z and pose using probabilistic quaternion fitting, falling into the group of learning-based methods. Volpe and Circi [66] on the other hand try to extract features using the KAZE algorithm [67], K-means clustering and RANSAC, and track them using the Kande-Lucas-Tomasi feature tracking algorithm [68]. Sharma and D'Amico [61] uses a similar feature detection for bounding box detection and solves pose as fitting wire-mesh model of the target in the bounding box problem. LSPNet [69] follows a similar approach but regresses pose a quaternion directly using the ROI's and bounding boxes generated. SU-Net [70] improves upon feature-learning methods by concatenating features using a residual-skip fusion architecture that they introduce as DR-U-Net, an improvement over the original U-net algorithm. All these algorithms are prone to problems faced by 2D vision based systems, auto-occlusions, lighting etc., [71] tries to overcome these challenges by introducing multi-dimensional loss function by integrating 2D-2D feature matching loss, 2D-3D key-point matching loss using a PnP solver and 3D-3D pose regression. This essentially removes any refinement modules like ICP, PnP or RANSAC as the pose is refined at 3 different stages i.e., initial pose estimate from 2D feature match loss which is refined using 3D key-point matching and finally direct numerical losses between poses. This allows for the network to improve its prediction on all 3 levels. Vela and Fasano [72] make use of feature detection algorithm to detect and localise arUco markers on the satellite and do pose estimation by doing a model reference matching treating the markers as key-points. Zhang and Hu [54] stand out as a pose estimation algorithm as it tries to exploit latent representation learning by comparing ground truth and observed representations of point clouds, to regress pose while updating the ground truth with each iteration to regress relative change in pose after initialisation. They use an ICP refinement module to improve accuracy of estimation after initial pose estimation by the algorithm.

The challenge for using machine/deep learning algorithms is their lack of flight worthiness i.e., some questions regarding AI that need to be answered regarding technical robustness, transparency and accountability of decision making, amongst many. For this reason researches like [60–62, 69, 70] become impractical because of their black-box like implementations, though allowing them high inference speeds which is an important factor. Some approaches fall short because of their computational or physical demand not being met by onboard systems or lower inference speeds like feature-based matching methods like [55, 72], which require special ArUco markers. A lot of 2D methodologies fall short in estimation accuracy as we discussed earlier that 3D information triumphs 2D as depth is difficult estimate by a monocular system. [54] allows for a new application using latent-learning for pose estimation, but

with their iteration-wise updated ground truth, it maybe difficult for the algorithm to deal with sudden uncertainties as it loses track of the original ground-truth information.

6 Conclusion

The pose estimation review over 2D and 3D methodologies reveals that pose derived from 3D information is always more accurate than that regressed from 2D information. This however does not overcome the advantages that camera systems have over LIDAR based systems in terms of reliability, cost, weight etc. The problems faced by 2D information systems remain constant as any disturbance in the image is a hindrance to the incoming information which is used to estimate pose. Some of these problems are occlusions, auto-occlusions, lighting conditions, unusual views, lack of annotated data. 2D methods still try to overcome these challenges by using feature-fusion [29], iterative-refinement [15], pose-tracking [19] etc. 3D methodologies on the other hand used to be computationally expensive and require an initial pose estimate which still doesn't guarantee fast or correct convergence of the pose. That is why most solutions use these as refinement modules on the last step. Under occlusion 2D methods currently perform better than 3D methods, as 3D methods do-not try to recover any lost information as they are not treated as complete information structures. With the recent growth in deep learning methodologies for 3D information like point clouds, it is established that point clouds can be reconstructed with as little as 5% of the complete point cloud available using latent representation learning[42–46, 54]. Inspired by these methodologies we could propose a new pose estimation methodology that leverages on latent-representation learning for direct pose estimation. Our idea is to translate the mathematics of rigid body transformations to latent representation state using Wasserstien Metrics [73], as Wasserstien Metrics are also known as optimal transport solutions which output the conversion of one distribution to another, working similar to rigid body transformations where our transformation matrix may act as a Wasserstien metric. This allows for the proposed algorithm to establish direct relationship between the latent representation state and inference level data (point cloud), making proposed algorithm flight-worthy, computationally inexpensive and accurate as Wasserstien Metrics are proved to be real distances or Euclidean distances. Leveraging on the point cloud completion thus, embedding completion methodology allows us to overcome problems that trouble vision based systems in the first place, i.e., occlusions, lighting, lack of data, etc.

Acknowledgements

This literature review is part of the ***PhD Project : Position and Attitude Estimation and Guidance for Space Rendezvous and Docking***, Siddharth Singh, under the supervision of Professors Hyo Sang Shin, Antonios Tsourdos and Leonard Felicetti. The goal of the PhD project being developing a monocular-vision based autonomous flight control system which can show case close proximity operations like docking, rendezvous, in-orbit facilities, de-orbiting out-of-service satellites among others. The current focus of the project is developing a robust SOTA localisation and navigation system, which later will be integrated with a control system. The project is funded by Centre for Autonomous and Cyberphysical Systems, SATM, Cranfield University, United Kingdom, MK43 0AL.

References

[1] Guoguang Du, Kai Wang, Shiguo Lian, and Kaiyong Zhao. Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review. *Artificial Intelligence Review*, 54, 2021. DOI: 10.1007/s10462-020-09888-5.

[2] Sabera Hoque, Shuxiang Xu, Ananda Maiti, Yuchen Wei, and Md. Yasir Arafat. Deep learning for 6D pose estimation of objects — A case study for autonomous driving. *Expert Systems with Applications*, 223:119838, 2023. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2023.119838.

[3] Ci-Jyun Liang, Kurt M. Lundeen, Wes McGee, Carol C. Menassa, SangHyun Lee, and Vineet R. Kamat. A vision-based marker-less pose estimation system for articulated construction robots. *Automation in Construction*, 104:80–94, 2019. ISSN: 0926-5805. DOI: https://doi.org/10.1016/j.autcon.2019.04.004.

[4] Dulari Bhatt, Chirag Patel, Hardik Talsania, Jigar Patel, Rasmika Vaghela, Sharnil Pandya, Kirit Modi, and Hemant Ghayvat. CNN Variants for Computer Vision: History, Architecture, Application, Challenges and Future Scope. *Electronics*, 10(20), 2021. ISSN: 2079-9292. DOI: 10.3390/electronics10202470.

[5] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes, 2018. `http://arxiv.org/abs/1711.00199`.

[6] Giorgia Marullo, Leonardo Tanzi, Pietro Piazzolla, and Enrico Vezzetti. 6D object position estimation from 2D images: a literature review. *Multimedia Tools and Applications*, 82, 2023. DOI: 10.1007/s11042-022-14213-z.

[7] Fang Wang and Zijian Zhao. A survey of iterative closest point algorithm. In *2017 Chinese Automation Congress (CAC)*, pages 4395–4399, 2017. DOI: 10.1109/CAC.2017.8243553.

[8] Martin A Fischler and Robert Coy 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. DOI: 10.1145/358669.358692.

[9] John A. Christian and Spott P. Cryan. A Survey of LIDAR Technology and Its Use in Spacecraft Relative Navigation. In *AIAA Guidance, Navigation and Control Conference*, 2014. DOI: 10.2514/6.2013-4641.

[10] Hansheng Chen, Wei Tian, Pichao Wang, Fan Wang, Lu Xiong, and Hao Li. EPro-PnP: Generalized End-to-End Probabilistic Perspective-n-Points for Monocular Object Pose Estimation, 2023.

[11] George Lentaris, Konstantinos Maragos, Ioannis Stratakos, Lazaros Papadopoulos, Odysseas Papanikolaou, Dimitrios Soudris, Manolis Lourakis, Xenophon Zabulis, David Gonzalez-Arjona, and Gianluca Furano. High-Performance Embedded Computing in Space: Evaluation of Platforms for Vision-Based Navigation. *Journal of Aerospace Information Systems*, 15(4):178–192, 2018. DOI: 10.2514/1.I010555.

[12] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. DPOD : 6D Pose Object Detector and Refiner. In *International Conference on Computer Vision (ICCV)*, October 2019.

[13] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. DeepIM: Deep iterative matching for 6d pose estimation. *International Journal of Computer Vision*, 128(3):657–678, nov 2019. DOI: 10.1007/s11263-019-01250-9.

[14] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. Implicit 3D Orientation Learning for 6D Object Detection from RGB Images, 2019.

[15] Yan Xu, Kwan-Yee Lin, Guofeng Zhang, Xiaogang Wang, and Hongsheng Li. RNNPose: Recurrent 6-DoF Object Pose Refinement With Robust Correspondence Field Estimation and Pose Optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14880–14890, June 2022.

[16] Fabian Manhardt, Gu Wang, Benjamin Busam, Manuel Nickel, Sven Meier, Luca Minciullo, Xiangyang Ji, and Nassir Navab. CPS++: Improving Class-level 6D Pose and Shape Estimation From Monocular Images With Self-Supervised Learning, 2020.

[17] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model Based Training, Detection and Pose Estimation of Texture-Less 3D Objects in Heavily Cluttered Scenes. In Kyoung Mu Lee, Yasuyuki Matsushita, James M. Rehg, and Zhanyi Hu, editors, *Computer Vision – ACCV 2012*, pages 548–562, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN: 978-3-642-37331-2.

[18] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6D Object Pose Estimation Using 3D Object Coordinates. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 536–551, Cham, 2014. Springer International Publishing. ISBN: 978-3-319-10605-2.

[19] Myung-Hwan Jeon and Ayoung Kim. PrimA6D: Rotational Primitive Reconstruction for Enhanced and Robust 6D Pose Estimation, 2020.

[20] Jin Liu and Sheng He. 6D Object Pose Estimation without PnP, 2019.

[21] Chen Song, Jiaru Song, and Qixing Huang. HybridPose: 6D Object Pose Estimation under Hybrid Representations. *CoRR*, 2020.

[22] Arul Selvam Periyasamy, Arash Amini, Vladimir Tsaturyan, and Sven Behnke. YOLOPose V2: Understanding and Improving Transformer-based 6D Pose Estimation, 2023.

[23] Thomas Jantos, Mohamed Amin Hamdad, Wolfgang Granig, Stephan Weiss, and Jan Steinbrener. PoET: Pose Estimation Transformer for Single-View, Multi-Object 6D Pose Estimation, 2022.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, 2015.

[25] Sukkeun Kim, Jeongho Kim, Jihoon Park, and Daewoo Lee. Vision-Based Pose Estimation of Fixed-Wing Aircraft Using You Only Look Once and Perspective-n-Points. *Journal of Aerospace Information Systems*, 18(9):659–664, 2021. DOI: 10.2514/1.I010975.

[26] Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement, 2018.

[27] Kiru Park, Timothy Patten, and Markus Vincze. Pix2Pose: Pixel-Wise Coordinate Regression of Objects for 6D Pose Estimation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. DOI: 10.1109/iccv.2019.00776, `https://doi.org/10.1109%2Ficcv.2019.00776`.

[28] Bugra Tekin, Sudipta N. Sinha, and Pascal Fua. Real-Time Seamless Single Shot 6D Object Pose Prediction, 2018.

[29] Yannick Bukschat and Marcus Vetter. EfficientPose: An efficient, accurate and scalable end-to-end 6D multi object pose estimation approach. *CoRR*, 2020.

[30] Mahdi Rad and Vincent Lepetit. BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects without Using Depth, 2018.

[31] Patrick Poirson, Phil Ammirato, Cheng-Yang Fu, Wei Liu, Jana Kosecká, and Alexander C. Berg. Fast Single Shot Detection and Pose Estimation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 676–684, 2016. DOI: 10.1109/3DV.2016.78.

[32] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". 2014.

[33] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, 2020.

[34] Jonathon Shlens. Notes on Kullback-Leibler Divergence and Likelihood, 2014.

[35] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion, 2019.

[36] Nuno Pereira and Luís A. Alexandre. MaskedFusion: Mask-based 6D Object Pose Estimation, 2020.

[37] Yisheng He, Haibin Huang, Haoqiang Fan, Qifeng Chen, and Jian Sun. FFB6D: A Full Flow Bidirectional Fusion Network for 6D Pose Estimation, 2021.

[38] Keunhong Park, Arsalan Mousavian, Yu Xiang, and Dieter Fox. LatentFusion: End-to-End Differentiable Reconstruction and Rendering for Unseen Object Pose Estimation, 2020.

[39] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-Squares Fitting of Two 3-D Point Sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9(5):698–700, 1987. DOI: 10.1109/TPAMI.1987.4767965.

[40] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun. PVN3D: A Deep Point-wise 3D Keypoints Voting Network for 6DoF Pose Estimation, 2020.

[41] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation, 2017.

[42] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R. Martin, and Shi-Min Hu. PCT: Point Cloud Transformer, 2020.

[43] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-BERT: Pre-training 3D Point Cloud Transformers with Masked Point Modeling, 2022.

[44] Xumin Yu, Yongming Rao, Ziyi Wang, Zuyan Liu, Jiwen Lu, and Jie Zhou. PoinTr: Diverse Point Cloud Completion with Geometry-Aware Transformers. In *ICCV*, 2021.

[45] Rahul Sajnani, Adrien Poulenard, Jivitesh Jain, Radhika Dua, Leonidas J. Guibas, and Srinath Sridhar. ConDor: Self-Supervised Canonicalization of 3D Pose for Partial Shapes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

[46] Ruiqi Wang, Xinggang Wang, Te Li, Rong Yang, Minhong Wan, and Wenyu Liu. Query6DoF: Learning Sparse Queries as Implicit Shape Prior for Category-Level 6DoF Pose Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14055–14064, October 2023.

[47] Ge Gao, Mikko Lauri, Xiaolin Hu, Jianwei Zhang, and Simone Frintrop. CloudAAE: Learning 6D Object Pose Regression with On-line Data Synthesis on Point Clouds, 2021.

[48] Vedant Bhandari, Tyson Govan Phillips, and Peter Ross McAree. Real-Time 6-DOF Pose Estimation of Known Geometries in Point Cloud Data. *Sensors*, 23(6), 2023. ISSN: 1424-8220. DOI: 10.3390/s23063085.

[49] Frederik Hagelskjær and Anders Glent Buch. Pointvotenet: Accurate Object Detection And 6 DOF Pose Estimation In Point Clouds. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2641–2645, 2020. DOI: 10.1109/ICIP40778.2020.9191119.

[50] Wei Chen, Jinming Duan, Hector Basevi, Hyung Jin Chang, and Ales Leonardis. PointPoseNet: Point Pose Network for Robust 6D Object Pose Estimation. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2813–2822, 2020. DOI: 10.1109/WACV45572.2020.9093272.

[51] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[52] Alec Radford and Karthik Narasimhan. Improving Language Understanding by Generative Pre-Training. 2018. `https://api.semanticscholar.org/CorpusID:49313245`.

[53] Jingnan Shi. Perspective-n-Point:P3P, 2022. `https://jingnanshi.com/blog/pnp_minimal.html`.

[54] Shaodong Zhang, Weiduo Hu, and Wulong Guo. 6-DoF Pose Estimation of Uncooperative Space Object Using Deep Learning with Point Cloud. In *2022 IEEE Aerospace Conference (AERO)*, pages 1–7, 2022. DOI: 10.1109/AERO53065.2022.9843444.

[55] Tristan Tzschichholz, Lei Ma, and Klaus Schilling. Model-based spacecraft pose estimation and motion prediction using a photonic mixer device camera. *Acta Astronautica*, 68(7):1156–1167, 2011. ISSN: 0094-5765. DOI: https://doi.org/10.1016/j.actaastro.2010.10.003.

[56] Giuseppe Napolano, Claudio Vela, Alessia Nocerino, Roberto Opromolla, and Michele Grassi. A multi-sensor optical relative navigation system for small satellite servicing. *Acta Astronautica*, 207:167–192, 2023. ISSN: 0094-5765. DOI: https://doi.org/10.1016/j.actaastro.2023.03.008.

[57] Lorenzo Pasqualetto Cassinis, Robert Fonod, and Eberhard Gill. Review of the robustness and applicability of monocular pose estimation systems for relative navigation with an uncooperative spacecraft. *Progress in Aerospace Sciences*, 110:100548, 2019. ISSN: 0376-0421. DOI: https://doi.org/10.1016/j.paerosci.2019.05.008.

[58] Vincenzo Pesce, Roberto Opromolla, Salvatore Sarno, Michèle Lavagna, and Michele Grassi. Autonomous relative navigation around uncooperative spacecraft based on a single camera. *Aerospace Science and Technology*, 84:1070–1080, 2019. ISSN: 1270-9638. DOI: https://doi.org/10.1016/j.ast.2018.11.042.

[59] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. EPnP: An accurate o(n) solution to the PnP problem. *Int. J. Comput. Vis.*, 81(2):155–166, Feb. 2009. DOI: https://doi.org/10.1007/s11263-008-0152-6.

[60] Thaweerath Phisannupawong, Patcharin Kamsing, Peerapong Torteeka, Sittiporn Channumsin, Utane Sawangwit, Warunyu Hematulin, Tanatthep Jarawan, Thanaporn Somjit, Soemsak Yooyen, Daniel Delahaye, and Pisit Boonsrimuang. Vision-Based Spacecraft Pose Estimation via a Deep Convolutional Neural Network for Noncooperative Docking Operations. *Aerospace*, 7(9), 2020. ISSN: 2226-4310. DOI: 10.3390/aerospace7090126.

[61] Sumant Sharma and Simone D'Amico. Neural Network-Based Pose Estimation for Noncooperative Spacecraft Rendezvous. *IEEE Transactions on Aerospace and Electronic Systems*, 56(6):4638–4658, 2020. DOI: 10.1109/TAES.2020.2999148.

[62] Xinghao Yang, Janmei Song, Haoping She, and Haichao Li. Pose estimation of non-cooperative spacecraft based on Convelutional Neural Network. In *2021 40th Chinese Control Conference (CCC)*, pages 8433–8438, 2021. DOI: 10.23919/CCC52363.2021.9549564.

[63] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions, 2014.

[64] Tae Ha Park, Marcus Martens, 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)*. IEEE, mar 2022. DOI: 10.1109/aero53065.2022.9843439, `https://doi.org/10.1109%2Faero53065.2022.9843439`.

[65] Pedro F. Proença and Yang Gao. Deep Learning for Spacecraft Pose Estimation from Photorealistic Rendering. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020. DOI: 10.1109/ICRA40945.2020.9197244, `https://ieeexplore.ieee.org/document/9197244`.

[66] Renato Volpe, Christian Circi, Marco Sabatini, and Giovanni B. Palmerini. GNC architecture for an optimal rendezvous to an uncooperative tumbling target using passive monocular camera. *Acta Astronautica*, 196:380–393, 2022. ISSN: 0094-5765. DOI: https://doi.org/10.1016/j.actaastro.2020.10.038.

[67] Pablo Fernández Alcantarilla, Adrien Bartoli, and Andrew J. Davison. Kaze features. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*, pages 214–227, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN: 978-3-642-33783-3.

[68] C Tomasi and T Kanade. Detection and tracking of point features (tech. rep. cmucs-91-132), 1991.

[69] Albert Garcia, Mohamed Adel Musallam, Vincent Gaudilliere, Enjie Ghorbel, Kassem Al Ismaeil, Marcos Perez, and Djamila Aouada. LSPnet: A 2D Localization-oriented Spacecraft Pose Estimation Neural Network, 2021.

[70] Hu Gao, Zhihui Li, Ning Wang, Jingfan Yang, and Depeng Dang. SU-Net: pose estimation network for non-cooperative spacecraft on-orbit. *Scientific Reports*, 13(1):11780, 2023.

[71] Taeyeop Lee, Byeong-Uk Lee, Inkyu Shin, Jaesung Choe, Ukcheol Shin, In So Kweon, and Kuk-Jin Yoon. UDA-COPE: Unsupervised Domain Adaptation for Category-level Object Pose Estimation, 2022.

[72] Claudio Vela, Giancarmine Fasano, and Roberto Opromolla. Pose determination of passively cooperative spacecraft in close proximity using a monocular camera and AruCo markers. *Acta Astronautica*, 201:22–38, 2022. ISSN: 0094-5765. DOI: https://doi.org/10.1016/j.actaastro.2022.08.024.

[73] Arijit Sehanobish, Neal Ravindra, and David van Dijk. Permutation invariant networks to learn Wasserstein metrics, 2021.