STPose: 6D object pose estimation network based on sparse attention and cross-layer connection

Shihao Chen[1]
2005120134@st.gxu.edu.cn
Xiaobing Li[1]
2004200318@st.gxu.edu.cn
Keduo Yan[1]
2003340203@st.gxu.edu.cn
Yong Li[1]†
yongli@gxu.edu.cn
Dongxu Gao[2]
Dongxu.Gao@port.ac.uk

[1] Guangxi Key Laboratory of Intelligent
Control and Maintenance of Power
Equipment
School of Electrical Engineering
Guangxi University
Nanning 530004, China.

[2] School of Computing
University of Portsmouth
Portsmouth PO1 3HE, UK

Abstract

The 6D object pose estimation technique provides accurate and rich coordinate information for robots to grasp target objects, while implementing the algorithms of this technique in industry often requires consideration of smaller cost loss. In this paper, we propose STPose, a transformer-based pose estimation network using only RGB images as input. Our network is based on PoET and proposes to reduce the computational parameters of the model with convergence efficiency by introducing a sparse attention method and an encoder cross-layer connection method. We also propose a system that enables easy and automatic implementation of labeled pose estimation datasets, since no research has been done to apply this technique to the power environment. Using this system, we produce a pose estimation dataset, the RCV dataset, targeting power device tools.STPose provides the best results among the currently studied algorithms on the RCV dataset and outperforms PoET (RGB-input-only Sota method) by 2.4% on the difficult YCB-V dataset. We also conduct an experimental analysis of the RCV dataset's features and difficulties. The project is available for public use at https://github.com/Agatha7k/STPose.

1 Introduction

With the advent of embodied intelligence and its subsequent growth, a growing number of researchers have expressed a strong interest in robot operation via autonomous planning[9, 16]. Among them, 6D pose estimation is a technology required to realize robot grasping operation, and robot grasping realization is an inevitable aspect of robot autonomous planning

†Corresponding author.

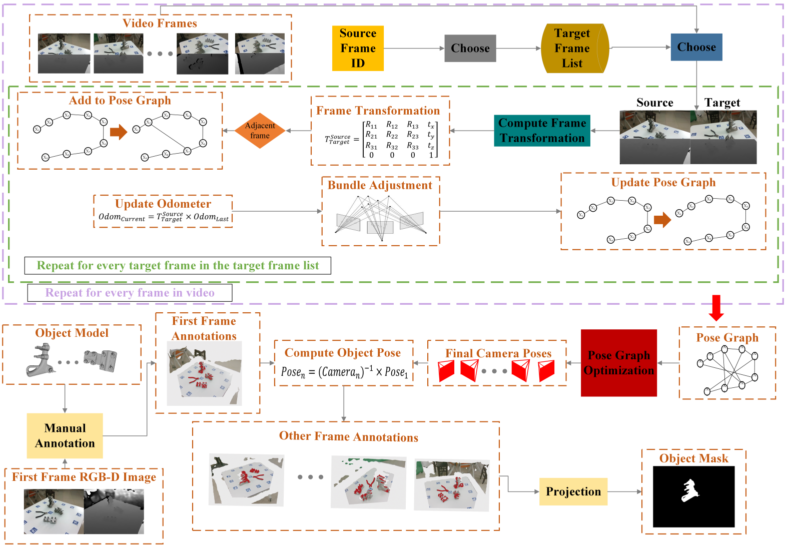
Video Frames
Source Frame ID
Choose
Target Frame List
Choose
Source
Target
Compute Frame Transformation
Frame Transformation
$T_{Target}^{Source} = \begin{bmatrix} R_{11} & R_{12} & R_{13} & t_x \\ R_{21} & R_{22} & R_{23} & t_y \\ R_{31} & R_{32} & R_{33} & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix}$
Adjacent frame
Add to Pose Graph
Update Odometer
$Odom_{Current} = T_{Target}^{Source} \times Odom_{Last}$
Bundle Adjustment
Update Pose Graph
Repeat for every target frame in the target frame list
Repeat for every frame in video
Pose Graph
Pose Graph Optimization
Final Camera Poses
Compute Object Pose
$Pose_n = (Camera_n)^{-1} \times Pose_1$
First Frame Annotations
Object Model
Manual Annotation
First Frame RGB-D Image
Other Frame Annotations
Projection
Object Mask

Figure 1: Overall block diagram of the principle of automatic labeling system: integrates front-end visual odometry and back-end nonlinear optimization.

operation. Moreover, researchers now have the chance to use deep learning for 6D location estimation of objects thanks to the ongoing development of deep learning-based approaches. Deep learning-based 6D pose estimation algorithms[17, 20, 28, 29, 31] exhibit greater applicability, robustness, and accuracy when compared to classical methods[8, 23]. The drawback of existing deep learning methods is that they can only improve bit-pose estimation accuracy by requiring a large number of annotated datasets. The difficulty of creating large-scale datasets is incredibly high since data annotation is intricate and time-consuming. Furthermore, the majority of the current datasets[5, 6, 13, 14] lack relevance to the electric power scenarios and application environments for power operation, belt power operation, and other scenarios of bit pose estimation dataset. These datasets primarily consist of objects commonly found in retail packages or household goods, such as the Open X-Embodiment dataset[24], which boasts a very large sample size and encompasses a wide range of objects proposed by the Open X-Embodiment Collaboration. There are currently several advanced techniques available for annotating datasets: Liu[21] et al. employed a set of task-specific model acquisition tools, but in order to complete this task, they had to rely on expensive experimental tools, which presented a challenge because they were unable to sample the dataset simultaneously and the expensive tools could not utilise all of the available time. We have developed a nearly fully automated approach for annotating datasets based on camera trajectory estimation as a solution to the aforementioned issues. This annotation tool is straightforward, quick, and inexpensive. A computer and an RGBD camera are the only capital and labor expenses required for this technology, which can significantly lessen annotators' workloads and increase the accuracy and efficiency of annotations. Using this combination of tools, we created the RCV dataset, which is the first BOP dataset for electric power background.

We also propose a 6D pose estimation network with only RGB images as inputs, taking into account the cost overhead under industrial conditions. We compare some common 6D object pose estimation algorithms on this dataset, like FFB6D[11] and densefusion[5], and perform benchmark tests, which demonstrate that our algorithms are able to achieve RGB-only input sota effect; some of the current state-of-the-art pose estimation algorithms on the RCV dataset achieve much lower scores in poor conditions than in normal conditions, and there is plenty of room for improvement; these results suggest that the dataset created by our set of annotation tools is appropriate to serve as a benchmark dataset for the pose estimation task.

Our primary research outcomes are listed below: (1) Our proposal is an annotation pipeline that utilizes camera trajectory estimation and includes both automatic and manual annotation techniques. (2) We produce a 6D pose estimation dataset applied to power scenarios using an automated labeling system.(3) We propose a transformer-based 6D pose estimation algorithm for RGB image inputs only and perform evaluation experiments on the YCB-V dataset and the RCV dataset.

2 Related Work

2.1 Labeling methods

6D pose estimate datasets are becoming more and more popular as technology advances[19, 27]. Six-degree-of-freedom pose is traditionally annotated manually, which is a time-consuming and error-pronemethod that involves matching a 3D object model with its representational properties in a 2D real image. The potential for automated 6D pose annotation is expanding along with the automation level. To create and annotate datasets, for instance, Baca[2] et al. created an almost entirely automated annotation technique. By combining evaluation measures and existing datasets, Hoden et al. created the Benchmarking Challenge for Six Degree of Freedom Pose Estimation Task[15] (BOP). Our approach has more advantages over Baca et al.'s almost fully automated approach to dataset generation and annotation, as well as Yuan[33] et al.'s motion sensor-based reconstruction of the object model. Specifically, we avoid the need for excessively complicated capturing devices and data acquisition tools, and the annotation process can be completed with just a computer and a D435i camera, which makes our annotation system more practical and effective.

2.2 Datasets

The two primary contexts covered by the datasets in the field of 6D pose estimation are home and industrial. The majority of the datasets in the home environment center on commonplace items or children's toys. For instance, the Linemod dataset[12] , which includes 13 low-textured objects in 13 video sequences without the use of generated images, was chosen for the home office setting. The difficulty is raised by the Linemod-Occluded dataset[13] , which includes additional difficulties such as truncation, object occlusion, and lighting fluctuations. On the other hand, one of the most difficult datasets currently accessible is the YCB-Video dataset[32], which is built in an office setting and has 92 video sequences and 21 items that also have illumination and occlusion problems. Thirty industrially important low-textured items, some of which are composites of other things and show some symmetry and shape and size similarity, are included in the T-LESS dataset[6].

One of the reasons we started our work was because we discovered that there was no pose estimate dataset available for the electrical industrial domain. The RCV dataset is designed to be trained and tested in a variety of lighting and occlusion scenarios, just like these traditional public datasets. Simultaneously, the RCV dataset must solve the simultaneous hurdles of classifying things with more reflective metals in a given viewing angle and detecting the presence of considerable occlusion for objects with varied volume sizes.

2.3 Methods

Unlike the previous models for pose estimation[11, 25, 28, 32], Trans6D[34] was the first study to use a Transformer model in the field of 6D pose estimation. The Transformer model was created in both pure and hybrid forms, and it was enhanced with a graphical convolutional network for the purpose of extracting local point cloud features and geometric sensing. These additions effectively imposed constraints on the Transformer model. A Transformer-based system was proposed by Amini[1] et al. that can regress the 6D pose of several objects from a single image. By adding translational and rotational heads to DETR[7], they were able to train the complete network end-to-end[7]. However, because this method relies on symmetric perceptual loss, it requires 3D models of the objects. In the meantime, PoET[18] presents a technique that can estimate the translation and rotation of an object in the camera coordinate system straight from a single RGB image, without the need for any extra data (such as a depth map, 3D model, or object symmetry information). PoET improves pose estimate accuracy by retaining the whole picture feature mapping and using the ROI as an additional input to the Transformer network, in contrast to previous methods that rely on ROIs for pose estimation.

Our network, which draws inspiration from PoET, incorporates the encoder cross-layer connection mechanism and sparse attention mechanism while preserving the PoET model's deformable convolution mechanism, which guarantees pose estimation accuracy while accelerating convergence and significantly enhancing the model's training efficiency, leading to more precise and useful outcomes in power industry scenarios.

3 Materials and Methods

3.1 Select Object Set

In order to enhance robot performance in identifying and grasping during electric power operations, our objective is to establish a benchmark dataset on electric power fixture devices. As a result, we decided to create this dataset using 16 items that are frequently seen in scenarios involving power operations. Taking into account the various physical characteristics of everyday and industrial devices, we selected two special cases of devices that have the same exterior texture and shape but different sizes (such as a high-voltage cable clamp or a vise with varying sizes) and the same shape and size but different exterior textures (such as diagonal jaw pliers with different colors). In these cases, it is challenging to accurately differentiate objects categories relying solely on the algorithms to accurately distinguish object classes and accurately estimate their poses. In order to properly take use of the complementarity between these two forms of information, the RCV dataset requires algorithms that can effectively fuse the geometric information with the texture information.

3.2 Making 3D physical models

After conducting numerous device scanning experiments, we determined that the EinScan Pro 2X 2020 handheld 3D scanner had the following restrictions and used it to create 3D models of the objects: Some very large devices cannot be scanned since the scanner cannot handle devices with lengths more than 500 cm. After scanning with this scanner, we have 16-ply files. Object visibility was 94.26% on average, with a wide range of orientations, and object distances ranging from 33.2 to 121.25 cm, with an average of 70.29 cm. There were nine things in total that were symmetrical and seven that were not.A schematic of every device in the RCV dataset is presented in Figure 2.

3.3 Capturing dataset images

We have an Intel Realsense d435i RGBD camera that we used to take RGB and depth pictures. The camera has a resolution of 1280*720 and can take both RGB and depth pictures.We separated the power devices into 80 training sets and 20 test sets, then used camera distances ranging from 0.5m to 1m to capture the device images under various lighting and occlusion situations. Using five different lighting conditions:(1) normal light with sparse distribution of devices (2) one-sided strong light exposure with sparse distribution of devices (3) dark light with sparse distribution of devices,(4) normal light with dense distribution of devices and (5) dark light with dense distribution of devices—each training set and test set was photographed. The test set only recorded 600 frames of video sequences, whereas each training set recorded 1800 frames in each condition. In the end, we were able to gather 144000 frames of video sequences, with 12,000 frames for the test set and 14,000 frames for the training set, where every device was allocated equally.

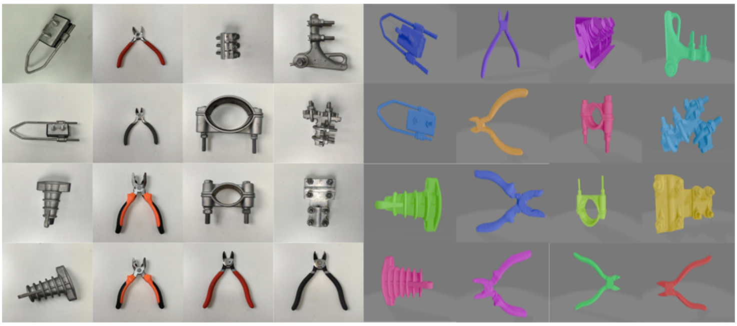

Figure 2: Demonstration of RCV dataset devices

3.4 Automated labeling of datasets

Using a set of video sequences as a unit, we apply the AruCo code fusion SIFT (Scale-invariant feature transform) algorithm[22] to extract the sparse feature points in the image. This involves fusing the local features extracted by the SIFT algorithm with the position and pose information of the AruCo code to create an all-encompassing feature descriptor. We calculate the point cloud surface overlap rate when generating the point cloud features, and use that information to filter and remove invisible objects in order to improve the accuracy of feature point matching. This method enhances the algorithm's robustness and accuracy by combining local and global data. After that, feature point matching is carried out in accordance with the features that have been described to obtain matching points. Lastly, the 3D to 3D relationship between feature points is used to solve the camera trajectory by the Iterative Closest Point (ICP) algorithm[3], which yields the transformation matrix between various camera coordinate systems. The corresponding camera pose matrices for each frame number are indicated as a_i, $i = 1,2,3,\ldots$ When $i = 1$s, it is the first frame of this video sequence. From equation (1), the $i+1$th frame position is calculated from the ith frame camera position.

$$a_{\text{target}} = R * a_{\text{source}} + t \tag{1}$$

In equation (1), a_{target} is the camera pose matrix of the target frame, a_{source} is the camera pose matrix of the current frame, R is the rotation matrix, and t is the translation matrix. By multiplying all of the transformation matrices from the first frame to the target number of frames by the pose of the first frame, this visual odometry method must determine the number of non-adjacent frames. This process often results in a loss of accuracy during multiplication, which can lead to errors. We present two back-end techniques, Pose Graph Optimization (PGO) and Bundle Adjustment (BA), to eliminate the aforementioned faults. While BA Optimization makes optimal adjustments to the camera pose with the spatial locations of feature points and eventually converges to the camera optical center to improve the accuracy of the camera pose estimation, Pose Graph Optimization can act on non-close frames to record all the sensory information accumulated during camera motion in a node-edge building manner to achieve global optimization of the pose. By comparing the projected value of the pixel coordinates with the measured value in terms of the pose error, or the reduced reprojection error, bit position optimization is carried out by utilizing the error. In terms of the metric, it can be stated as follows:

$$\min_{R_i,t_i} \sum_{i,j} \sigma_{ij} \|u_{ij} - v_{ij}\|_2 \tag{2}$$

Among them, R_i and t_i represent the rotation matrix and translation matrix of the camera frame P_i, u_{ij} represents the coordinates of the map point X_j projected onto the camera frame P_i, and v_{ij} represents the coordinates of the map point X_j reprojected onto the camera frame P_i. When the map point X_j is projected in the camera frame P_i, $\sigma_{ij} = 1$; otherwise, $\sigma_{ij} = 0$.

In Figure 1, the green frame portion is shown as follows: the pose transformation matrix with respect to the source frames is calculated for each target frame in the target frame list once a list of source and target frames has been chosen, and the odometry is updated to constantly optimize the pose map. During this time, BA optimization also helps to lower the error. The automatic annotation principle, which the purple frame portion is shown as follows: enter the green frame portion and continue in this manner until you have traversed every target frame list for this source frame. This process is done after selecting the list of target frames based on the chosen source frame ID.

3.5 Proposed STPose algorithmic framework

We present a deep learning-based object 6D pose estimation algorithm that can reliably and accurately estimate the target object's rotation and translation state in complex environmental conditions. It also exhibits strong robustness to changes in ambient illumination and can reliably and accurately estimate the target object's pose in situations where the object is severely occlusion and has weakly textured or untextured features on its surface. Impact. Subsequently, we will talk about the network's overall architecture and the implementation's unique enhancement.

The network's general layout is depicted in Figure 3. Our network is primarily divided into three sections:(1) Backbone feature extraction network: the input picture passes through this portion of the network to obtain the multi-scale feature map and predicted bounding box; this portion of the network can be substituted with any object target detector.(2) The transformer network, which is part of the network, is based on the PoET and introduces deformable convolution, encoder token sparsification, and encoder cross-layer connection mechanism. The output embedding is fed into the decoder along with the camera center coordinate information and the object query embedding filtered by the scoring network, so that the decoder's output embedding includes both the local and global information; this is done after receiving the multiscale features and bounding box obtained from the previous part of the network and feeding this information to the encoder.(3) Translation head and rotation head networks: Using the data from the previous network, this portion of the network enables us to estimate various bit pose information and category information of multiple objects at once, resulting in precise 6D bit pose estimation of objects.

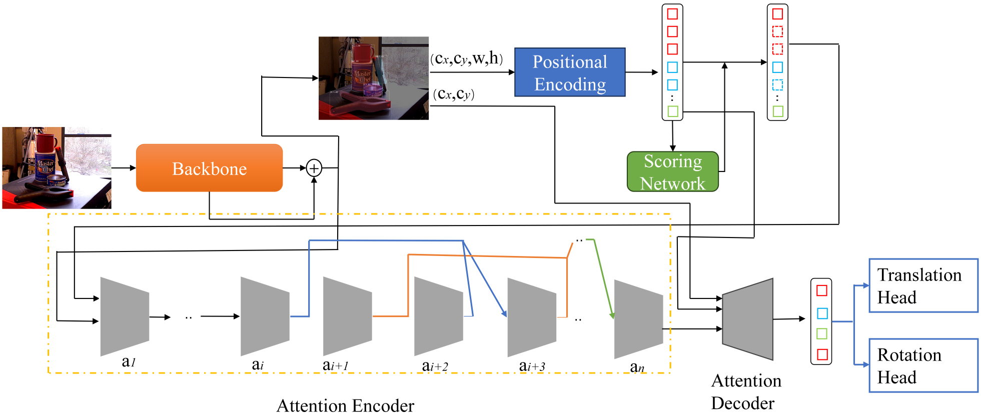

Figure 3: Overview diagram of the STPose algorithm

3.5.1 Transformer attention sparsification

The attention sparsification module of our STPose network is shown in the "scoring network" section of Figure 3. The feature map is x before entering the scoring network, and the scoring network is assumed to be s, which serves for measuring the significance of the tokens in x. We define a top-$k\%$ region X_k which contains the tokens in the top $k\%$ of significance. The token of the encoder i is updated as follows:

$$\mathbf{x}_i^j = \begin{cases} \mathrm{LN}\left(\mathrm{FFN}\left(\mathbf{z}_i^j\right) + \mathbf{z}_i^j\right) & j \in X_k, \text{ where } \mathbf{z}_i^j = \mathrm{LN}\left(\mathrm{DA}\left(\mathbf{x}_{i-1}^j, \mathbf{x}_{i-1}\right) + \mathbf{x}_{i-1}^j\right) \\ \mathbf{x}_{i-1}^j\Big) & j \notin X_k \end{cases} \tag{3}$$

where FFN is the feed-forward network, LN is layer normalization, and DA is deformable attention. To ensure that the information is not lost, lower the computational cost, and achieve the effect of encoder token sparsification, the information is passed to the token in the region even if it is not in the designated region.

3.5.2 Encoder cross-layer connection

Jump connection, such as the usage of long program connectivity in U-Net[26] and short program connectivity in ResNet[10], is a popular technique for enhancing the performance of deep learning networks. On the other hand, short-program jump connections may result in information loss due to poor memory, while long-program jump connections may result in inadequate fusion information representation because of large disparities in information features before and after. As a result, we decide on a medium-program jump link.

The encoder cross-layer connection graph is also displayed in Figure 3. The fusion information of encoders a_i and a_{i+2} is the input information of encoder a_{i+3}, assuming that the i-th encoder is a_i. In other words, the combination of the output data from encoder a_i and encoder a_{i+2} yields all of the encoder information, except the first and second encoder.

4 EXPERIMENTS

To evaluate the performance of our RCV dataset and validate our algorithm STPose, we run tests of STPose on the YCB-V dataset and experiment with several 6D pose estimation strategies on the RCV dataset. The RCV dataset comprises 3,505 frames from the training set divided into the validation set, 12,000 frames from the test set, and 144,000 frames from the training set. It is divided into different occlusion and illumination settings for group testing.

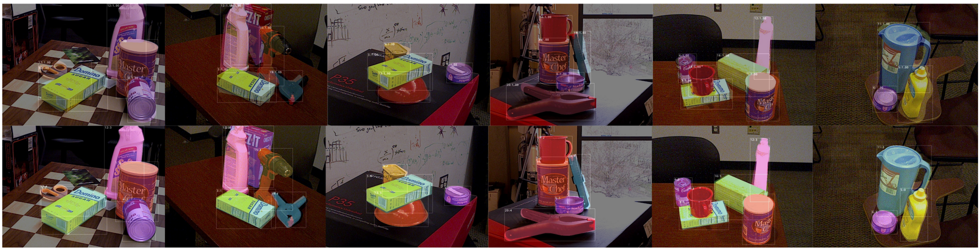

Figure 4: Figure 4: Qualitative results of STPose's predictions of relative 6D poses for the YCB-V dataset(Estimated poses in the top row, labeled poses in the bottom row)

4.1 Evaluation metrics

We use two metrics, Average Distance (ADD) and Average Distance to Nearest Point (ADD-S), to assess the bit pose estimation accuracy. where symmetric items are evaluated using ADD-S and asymmetric objects using ADD.

4.2 Evaluation on YCB-V Dataset

Our method optimizes the network parameters during training by using the Adam optimizer, which is based on the Pyorch framework implementation. Every experiment in this paper is performed on a desktop computer that has two NVIDIA RTX 3090 GPUs and an Intel® Xeon® E5-2680 v4 CPU.

Method	SilhoNet[4]	MCN[26]	T6D[1]	GDR-Net[30]	DeepIM[20]	PoET[18]	Our STPose
3D Model	Input + Sym	2D	Loss	PnP	IR	2D	2D
002 master chef can	83.6	91.2	91.9	**96.6**	93.1	92.9	91.8
003 cracker box	88.4	78.5	86.6	84.9	**91**	90.4	79.1
004 sugar box	88.8	85.1	90.3	98.3	96.2	94.5	**100**
005 tomato soup can	89.4	93.3	88.9	**96.1**	92.4	94	93.4
006 mustard bottle	91	91.9	94.7	**99.5**	95.1	94.8	99.7
007 tuna fish can	89.9	95.2	92.2	95.1	96.1	94	**100**
008 pudding box	89.1	84.9	85.1	94.8	90.7	93.8	**100**
009 gelatin box	94.6	92.1	86.9	95.3	94.3	92.7	**99.8**
010 potted meat can	84.8	90.8	83.5	82.9	86.4	**94.1**	93.5
011 banana	88.7	70	93.8	**96**	72.3	94.3	61
019 pitcher base	91.8	91.1	92.3	**98.8**	94.6	94.3	92.4
021 bleach cleanser	72	86.8	83	94.4	90.3	92.6	**99.5**
024 bowl	72.5	85	91.6	84	81.4	92.1	**99.1**
025 mug	92.1	91.9	89.8	96.9	91.3	94.1	**100**
035 power drill	82.9	87.2	88.8	91.9	92.3	**94.3**	93.8
036 wood block	79.2	87.2	90.7	77.3	81.9	92	**100**
037 scissors	78.3	80.2	83	68.4	75.4	92.5	**99.8**
040 large marker	83.1	66.4	74.9	87.4	86.2	81.6	**96.3**
051 large clamp	84.5	86.5	78.3	69.3	74.3	95.7	**100**
052 extra large clamp	88.4	79.5	54.7	73.6	73.2	96	**100**
061 foam brick	88.4	79.2	89.9	90.4	81.9	89.7	**100**
MEAN	85.8	86.9	86.2	89.1	88.1	92.8	**95.2**

Table 1: Quantitative evaluation results using the ADD-(S) metric on the YCB-V dataset for the entire test set, where data shown in bold are the highest scores among the different methods.

The YCB-V public dataset is used as a benchmark in the tests in this section to compare our suggested network STPose to other techniques that also only accept RGB inputs. Table 1 presents the evaluation results of STPose and other advanced RGB algorithms on YCB-V. The findings indicate that the bit-pose estimate accuracy of existing advanced RGB methods is not as high as our method. In order to decrease the regression prediction of rotations, a system similar to SilhoNet feeds the network symmetry information of 3D objects. This indicates that STPose has the potent capacity to model the global context of the image because it incorporates a transformer structure and does not require the input of extra data. As a pose estimation network that also uses a transformer structure, STPose is also 2.4% higher than the state-of-the-art (sota) algorithm PoET. One advantage of STPose is that it can reduce the amount of computation in attention by allowing the encoder token to be sparse. This is achieved by building a scoring network that chooses keys with strong representativeness, which lowers the number of keys input into the encoder of transformer number. This greatly reduces the computational complexity of the attention and can facilitate network convergence.

4.3 Ablation experiments

The ablation experiments are performed on the RCV dataset as a baseline and on the various mechanisms applied in this paper, and the experimental results obtained are presented in the following table.

In the above table, model (a) and model (b) are the PoET network and PoET network with applying deformable convolution mechanism, respectively, and model (c) and model

Model	(a)	(b)	(c)	(d)
DCN	✗	✓	✓	✓
TAS	✗	✗	✓	✓
ECN	✗	✗	✗	✓
Accuracy (%)	86.88	87.29	89.25	**89.46**

Table 2: Results of ablation experiments. "DCN" denotes deformable convolution, "TAS" denotes transformer attention sparsification, and "ECN" denotes encoder cross-layer connection.

(d) are the STPose network after incorporating the sparse attention mechanism and after incorporating the sparse attention mechanism as well as the cross-layer connection mechanism, respectively. As can be seen in Table 2, after incorporating the deformable convolution mechanism, the pose estimation accuracy of model (b) increases from 86.88% to 87.29%, which is an improvement of 0.41%, which indicates that the deformable convolution can be adjusted according to different shapes of objects, and improves the ability to understand the objects in the complex scene and the pose estimation relative to the traditional convolution; after incorporating the sparse attention mechanism, the accuracy of pose estimation of model (c) is improved from 87.29% to 89.25%, which is an improvement of 1.96%, indicating that the proposal of sparse attention mechanism greatly reduces the complexity of attention computation in the transformer network, and improves the computational performance and generalization ability of the STPose network; after adding the cross-layer connectivity mechanism, the accuracy of pose estimation of model (d) is improved from 89.25% to 89.46%, an improvement of 0.21%, which shows that the cross-layer connection mechanism can make the STPose network information difficult to be lost in the case of reaching at deeper encoder, making its convergence ability stronger and the pose estimation ability more accurate.

5 CONCLUSION

We create an object pose estimation dataset for power-operated devices, called the RCV dataset, using an automatic labeling approach for object 6D pose estimation dataset that only needs to name the first frame of a video sequence. This dataset encourages the use of object 6D location estimation algorithms in electric power scenes as it is the first publicly available dataset for electric power devices in this field. And our proposed STPose network outperforms current RGB methods.The encoder is the focus of our network's improvement; in fact, we have also tried to improve the structure of the decoder, such as exchanging the TRANSFORMER decoder self-attention layer with the cross-attention layer. Finally, we hope that it can be served as a useful and trustworthy link between computer vision and the power sector.

References

[1] Arash Amini, Arul Selvam Periyasamy, and Sven Behnke. T6d-direct: Transformers for multi-object 6d pose direct regression. In *DAGM German Conference on Pattern Recognition*, pages 530–544. Springer, 2021.

[2] Javier Gibran Apud Baca, Thomas Jantos, Mario Theuermann, Mohamed Amin Ham-

dad, Jan Steinbrener, Stephan Weiss, Alexander Almer, and Roland Perko. Automated data annotation for 6-dof ai-based navigation algorithm development. *Journal of Imaging*, 7(11), 2021.

[3] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. Spie, 1992.

[4] Gideon Billings and Matthew Johnson-Roberson. Silhonet: An rgb method for 6d object pose estimation. *IEEE Robotics and Automation Letters*, 4(4):3727–3734, 2019.

[5] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*, pages 536–551. Springer, 2014.

[6] Berk Calli, Aaron Walsman, Arjun Singh, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. Benchmarking in manipulation research: Using the yale-cmu-berkeley object and model set. *IEEE Robotics & Automation Magazine*, 22(3):36–52, 2015.

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[8] Mark Fiala. Artag, a fiducial marker system using digital techniques. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 590–596. IEEE, 2005.

[9] Agrim Gupta, Silvio Savarese, Surya Ganguli, and Li Fei-Fei. Embodied intelligence via learning and evolution. *Nature communications*, 12(1):5721, 2021.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Yisheng He, Haibin Huang, Haoqiang Fan, Qifeng Chen, and Jian Sun. Ffb6d: A full flow bidirectional fusion network for 6d pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3003–3013, 2021.

[12] Stefan Hinterstoisser, Stefan Holzer, Cedric Cagniart, Slobodan Ilic, Kurt Konolige, Nassir Navab, and Vincent Lepetit. Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In *2011 international conference on computer vision*, pages 858–865. IEEE, 2011.

[13] 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 *Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part I 11*, pages 548–562. Springer, 2013.

[14] Tomáš Hodan, Pavel Haluza, Štepán Obdržálek, Jiri Matas, Manolis Lourakis, and Xenophon Zabulis. T-less: An rgb-d dataset for 6d pose estimation of texture-less objects. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 880–888. IEEE, 2017.

[15] Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders GlentBuch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, et al. Bop: Benchmark for 6d object pose estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 19–34, 2018.

[16] David Howard, Agoston E Eiben, Danielle Frances Kennedy, Jean-Baptiste Mouret, Philip Valencia, and Dave Winkler. Evolving embodied intelligence from materials to machines. *Nature Machine Intelligence*, 1(1):12–19, 2019.

[17] Yinlin Hu, Joachim Hugonot, Pascal Fua, and Mathieu Salzmann. Segmentation-driven 6d object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3385–3394, 2019.

[18] Thomas Georg Jantos, Mohamed Amin Hamdad, Wolfgang Granig, Stephan Weiss, and Jan Steinbrener. Poet: Pose estimation transformer for single-view, multi-object 6d pose estimation. In *Conference on Robot Learning*, pages 1060–1070. PMLR, 2023.

[19] Josip Josifovski, Matthias Kerzel, Christoph Pregizer, Lukas Posniak, and Stefan Wermter. Object detection and pose estimation based on convolutional neural networks trained with synthetic data. In *2018 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 6269–6276. IEEE, 2018.

[20] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep iterative matching for 6d pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 683–698, 2018.

[21] Liu Liu, Wenqiang Xu, Haoyuan Fu, Sucheng Qian, Qiaojun Yu, Yang Han, and Cewu Lu. Akb-48: A real-world articulated object knowledge base. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14809–14818, 2022.

[22] David G Low. Distinctive image features from scale-invariant keypoints. *Journal of Computer Vision*, 60(2):91–110, 2004.

[23] Edwin Olson. Apriltag: A robust and flexible visual fiducial system. In *2011 IEEE international conference on robotics and automation*, pages 3400–3407. IEEE, 2011.

[24] Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.

[25] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[27] Stefan Thalhammer, Timothy Patten, and Markus Vincze. Sydpose: Object detection and pose estimation in cluttered real-world depth images trained using only synthetic data. In *2019 International Conference on 3D Vision (3DV)*, pages 106–115. IEEE, 2019.

[28] 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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3343–3352, 2019.

[29] Chen Wang, Roberto Martín-Martín, Danfei Xu, Jun Lv, Cewu Lu, Li Fei-Fei, Silvio Savarese, and Yuke Zhu. 6-pack: Category-level 6d pose tracker with anchor-based keypoints. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10059–10066. IEEE, 2020.

[30] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16611–16621, 2021.

[31] Bowen Wen, Chaitanya Mitash, Baozhang Ren, and Kostas E Bekris. se (3)-tracknet: Data-driven 6d pose tracking by calibrating image residuals in synthetic domains. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10367–10373. IEEE, 2020.

[32] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017.

[33] Honglin Yuan, Tim Hoogenkamp, and Remco C. Veltkamp. Robotp: A benchmark dataset for 6d object pose estimation. *Sensors*, 21(4), 2021. ISSN 1424-8220. doi: 10.3390/s21041299.

[34] Zhongqun Zhang, Wei Chen, Linfang Zheng, Aleš Leonardis, and Hyung Jin Chang. Trans6d: Transformer-based 6d object pose estimation and refinement. In *European Conference on Computer Vision*, pages 112–128. Springer, 2022.