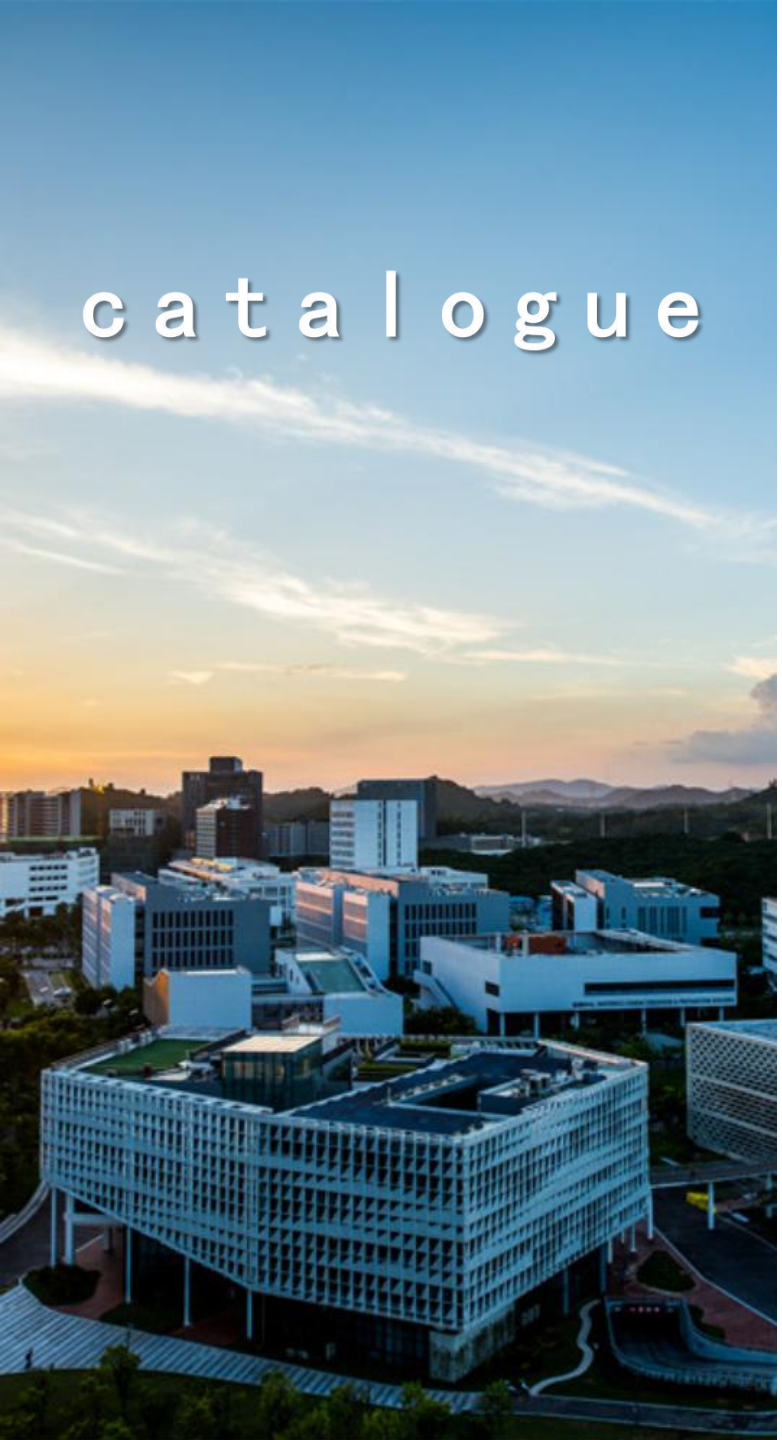
TLCD: A Transformer based Loop Closure Detection for Robotic Visual SLAM

Chenghao Li, Hongwei Ren, Minjie Bi, Chenchen Ding, Wenjie Li, Rumin Zhang, Xiaoguang Liu and Hao Yu

Speaker：Chenghao Li

Academic advisor: Hao Yu

This work was supported by the National Natural Science Foundation of China (NSFC) (Key Program Grant No. 62034007), Innovative Team Program of Education Department of Guangdong Province (Grant No. 2018KCXTD028), and Shenzhen Science and Technology Program (Grant No. KQTD20200820113051096) and NSQKJJ (Grant No. K21799101).



1 Introduction

2 Transformer Based Loop Closure Detection (TLCD)

3 Experiments and Results

4 Conclusion

1 Introduction

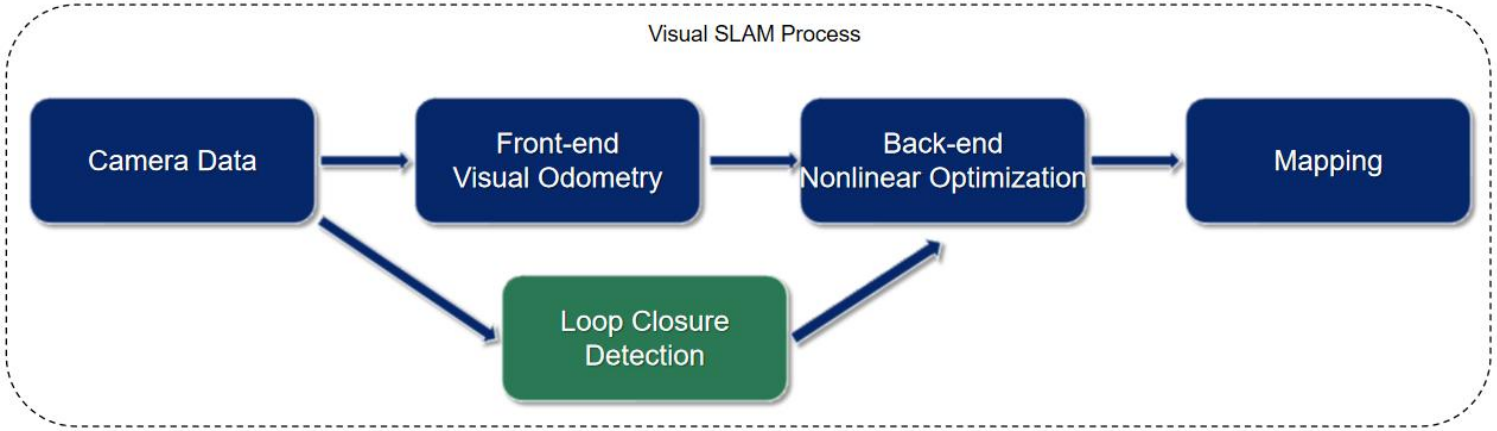
1 Introduction

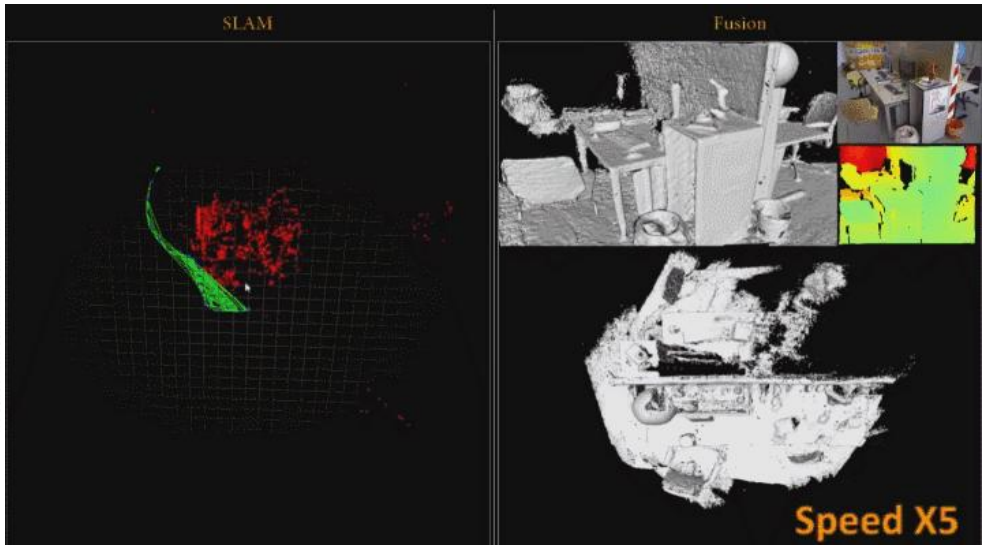
SLAM

Where I am? → Pose estimation

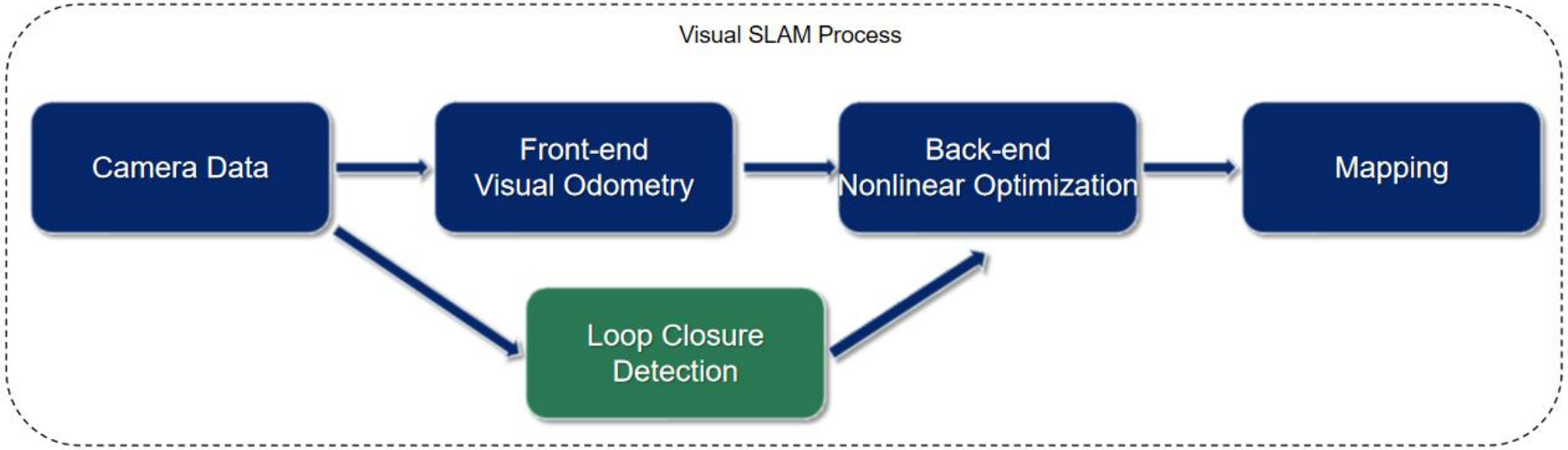
What is the environment around me? → Environment mapping

Simultaneous Localization and Mapping (SLAM)





1 Introduction



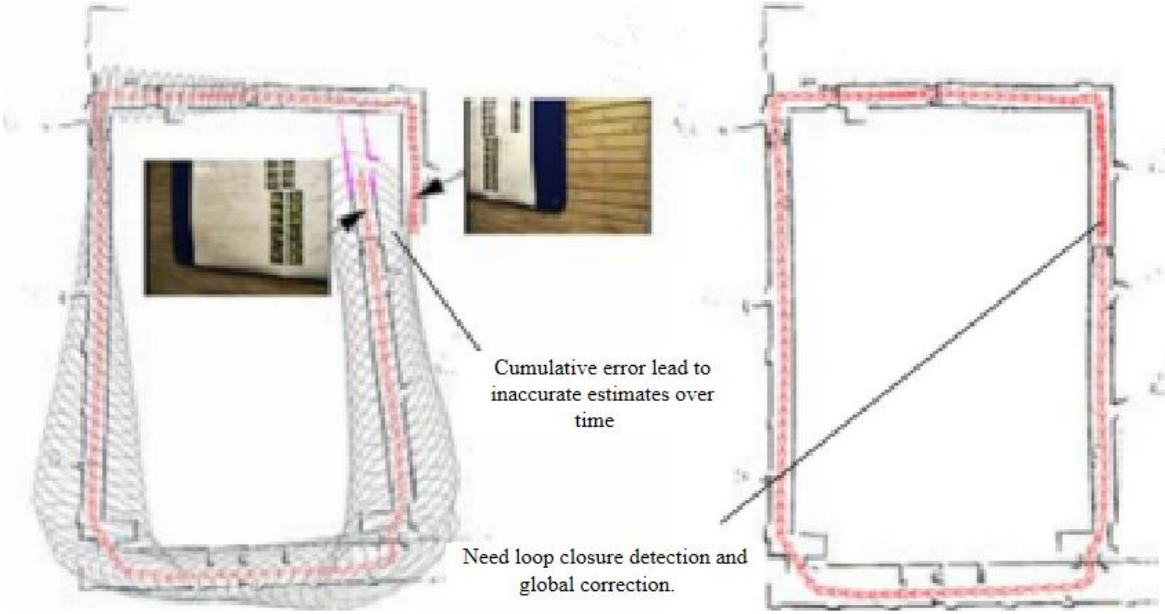
Visual Odometry (VO): Estimate the camera pose changes between adjacent sampling images, so as to estimate the motion trajectory ;

Back-end Optimization: The global trajectory and map are obtained by combining and optimizing the camera trajectory obtained by VO and loop closure detection information;

Loop Closure Detection (LCD): Determine whether the camera has a closed-loop trajectory, that is, determine whether the camera has passed through the same location;

Mapping: Build a map based on the trajectory.

Loop Closure Detection (LCD)

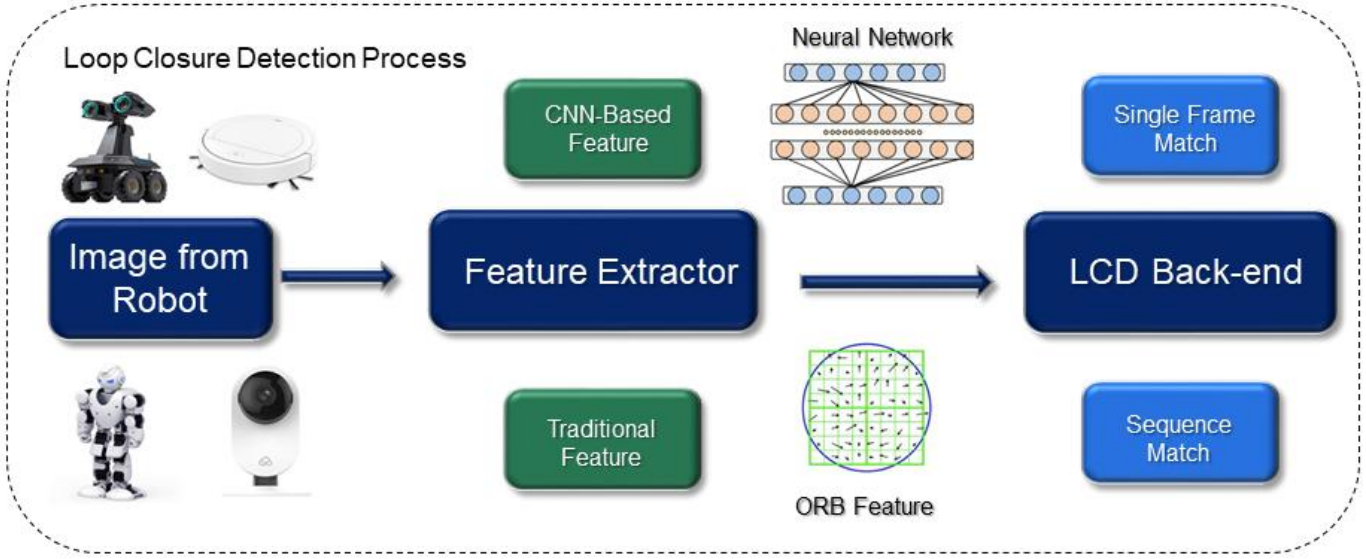


Cumulative Error Lead to Inaccurate Estimates

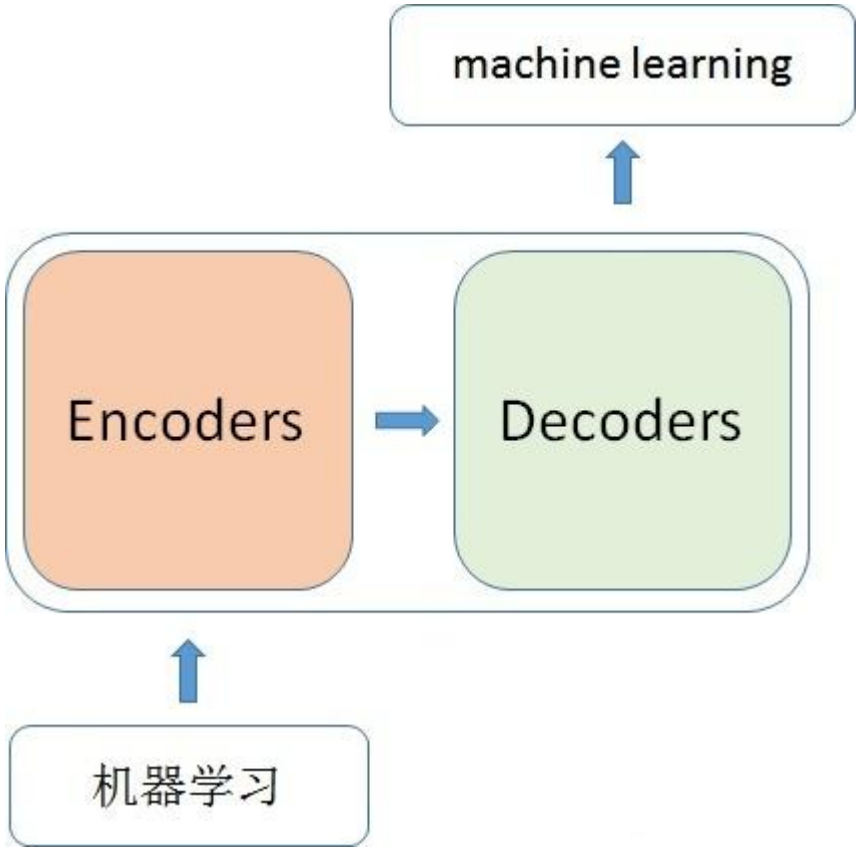
LCD aims to eliminate such cumulative errors as much as possible and make the trajectory that should be closed accurately.

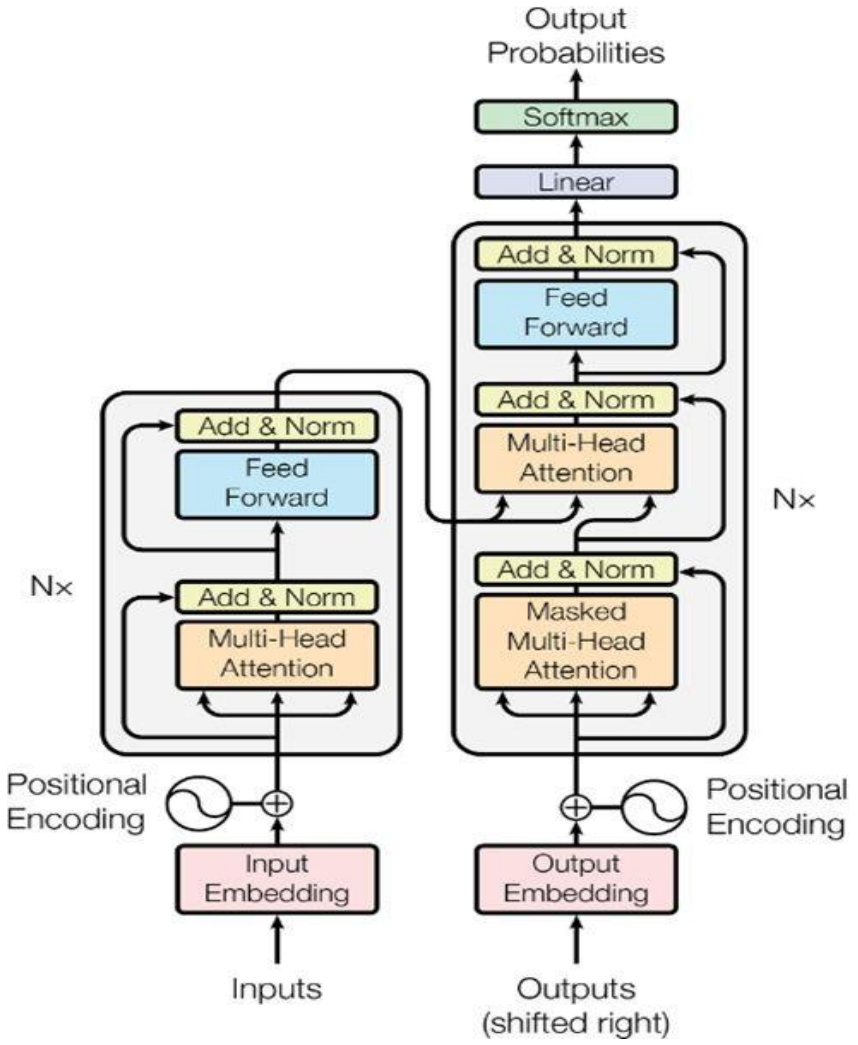
Modules for Different LCD Algorithms

Traditional Methods: SIFT, SURF, BoW, ORB, BRISK, etc.
CNN-based: PCANet, AlexNet, GoogLeNet, etc.



Transformer

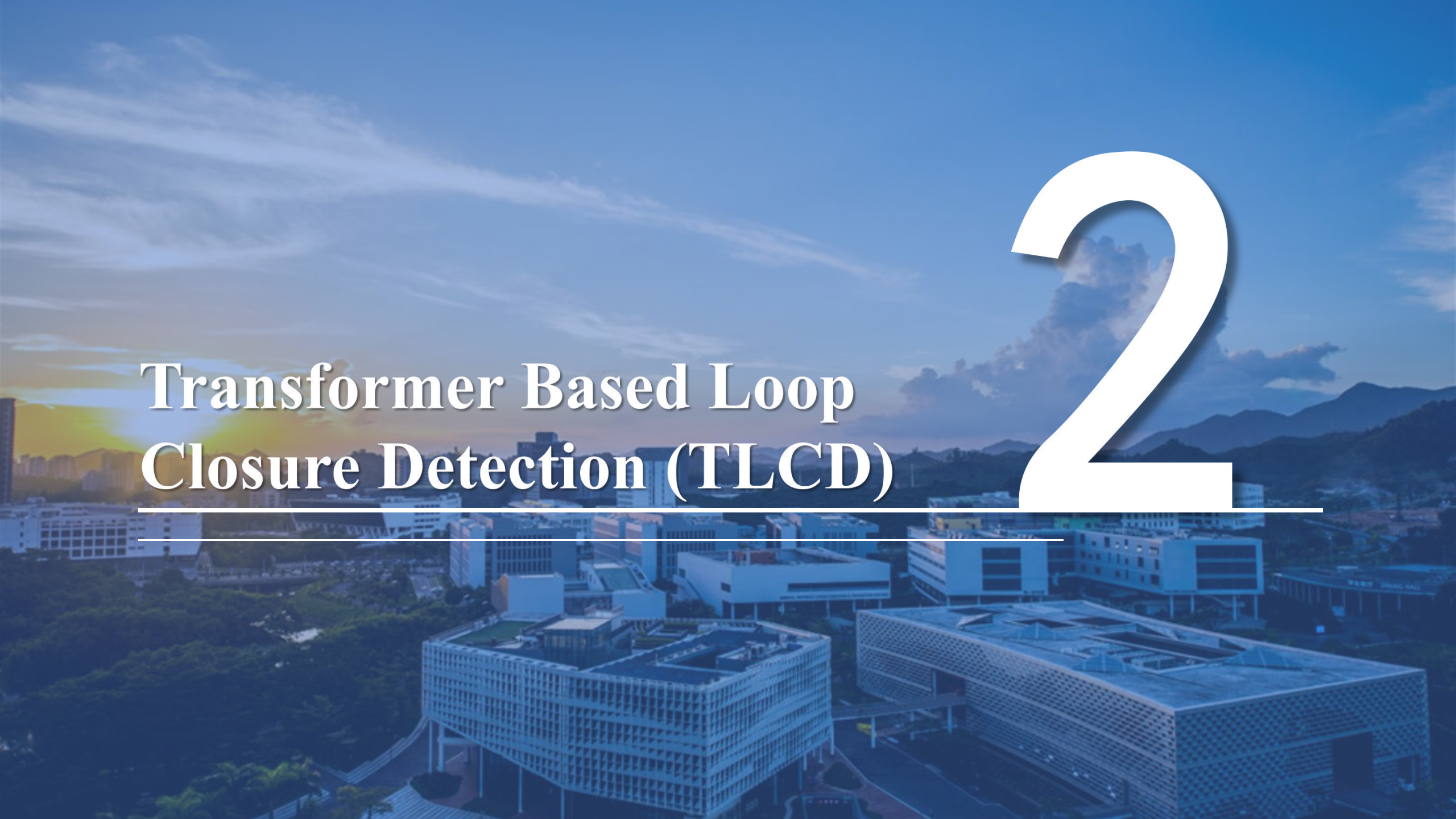




Contribution

This research proposes a vision transformer and sequence matching loop closure detection system called a transformer based loop closure detection for visual SLAM (TLCD). The most major contributions of this work are as follows:

1. **Feature extraction** of RGB images with **transformer**, generating image features required by the back-end of loop closure detection.
2. **Training transformer** using Places365 to get the weight of the scene classification.
3. **Sequence matching** is designed at the back-end of loop closure detection for improving the average accuracy.
4. The loop closure detection proposed by us achieves **competitive average accuracy**.

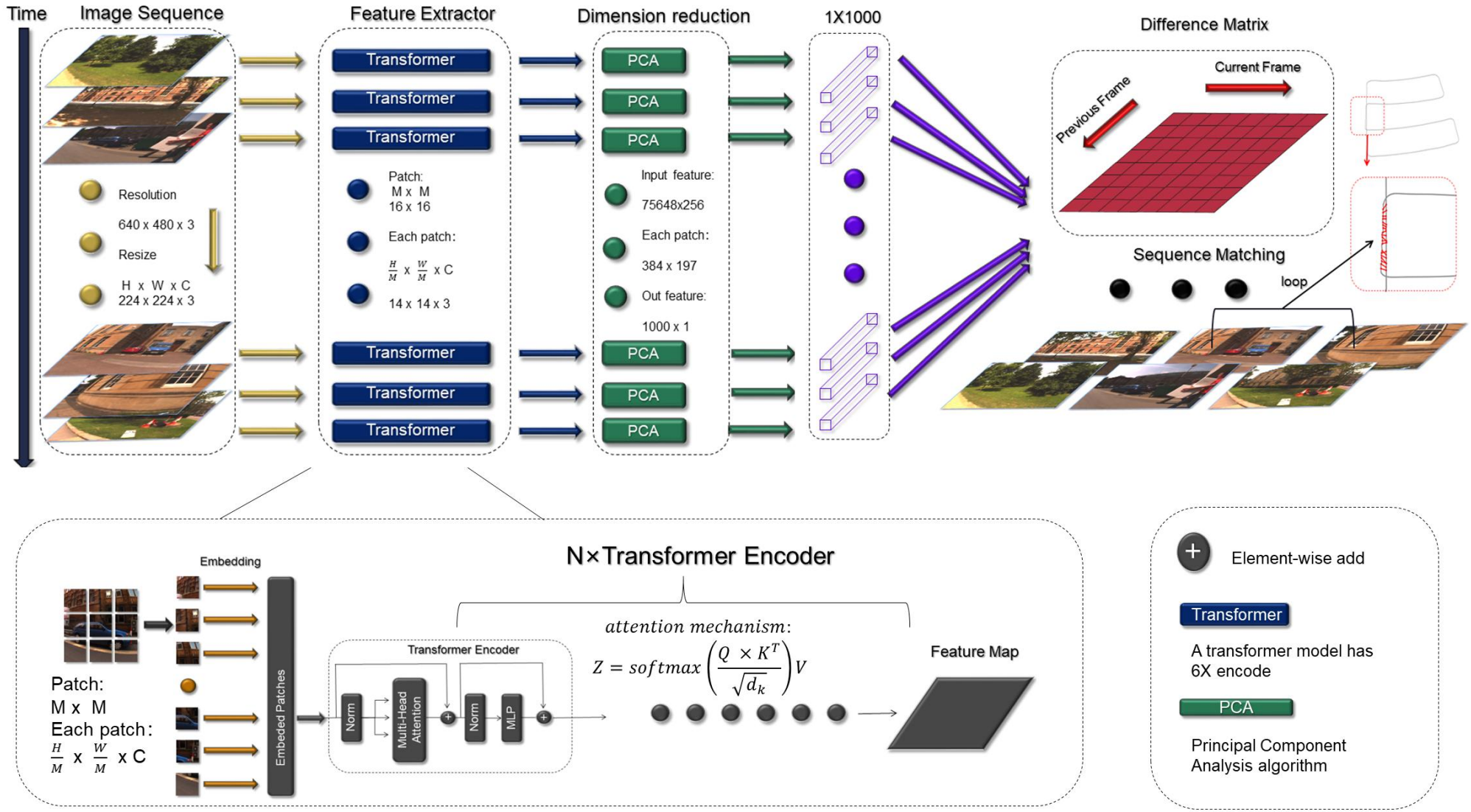


Transformer Based Loop Closure Detection (TLCD)

2

2 Transformer Based Loop Closure Detection (TLCD)

System Framework



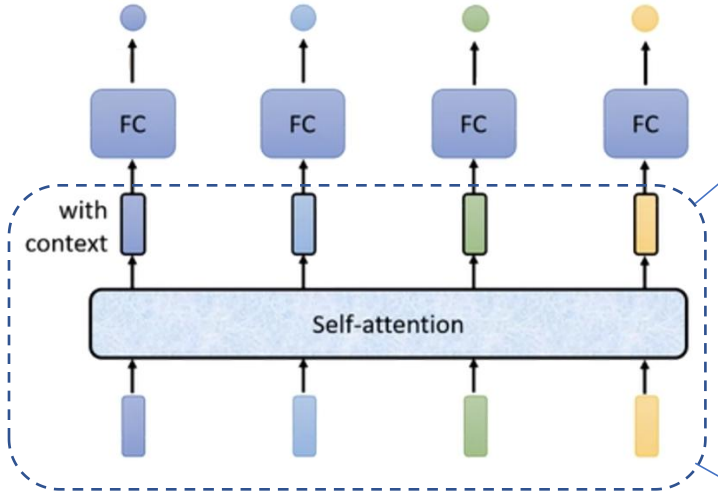
2 Transformer Based Loop Closure Detection (TLCD)

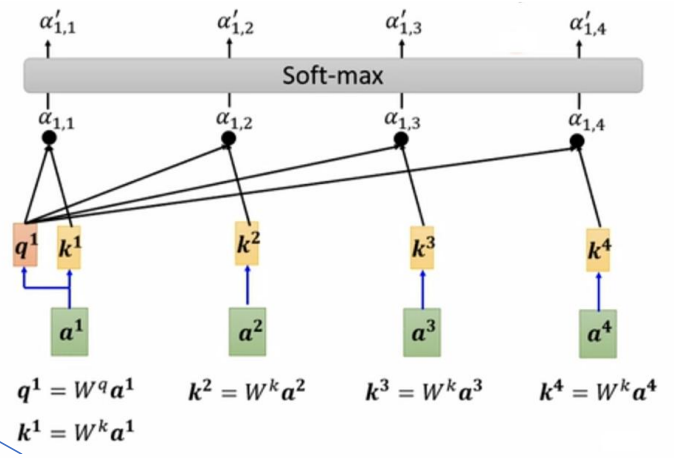
Feature Extractor Based on Transformer

$$Q = W^q I$$
$$K = W^k I$$
$$V = W^v I$$

Need learn

Self-attention

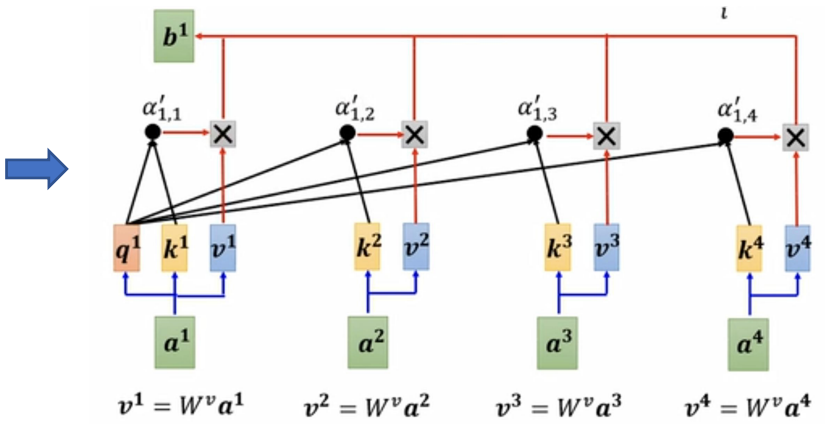




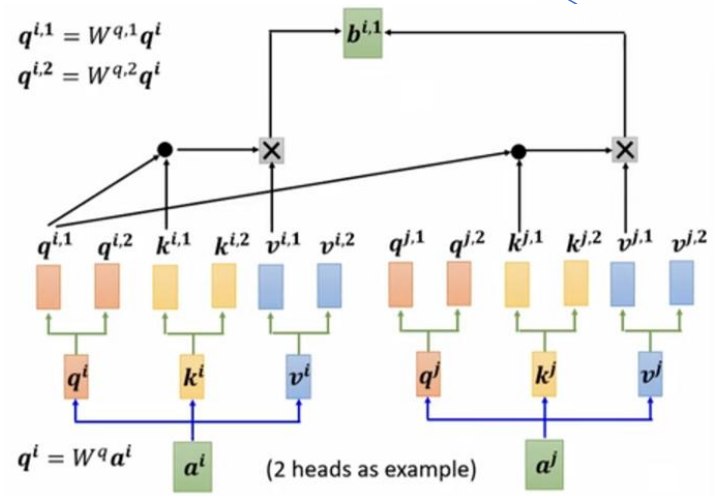
$$A' = softmax(A) = softmax(K^T Q)$$

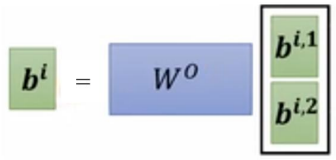
Attention Matrix

$$O = VA'$$

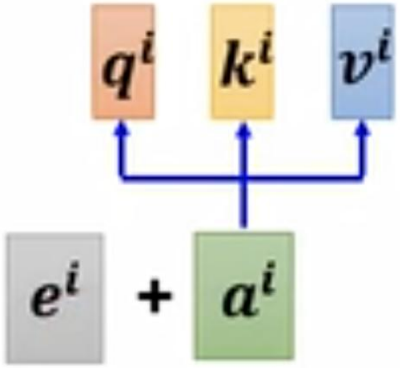


Multi-head Self-attention



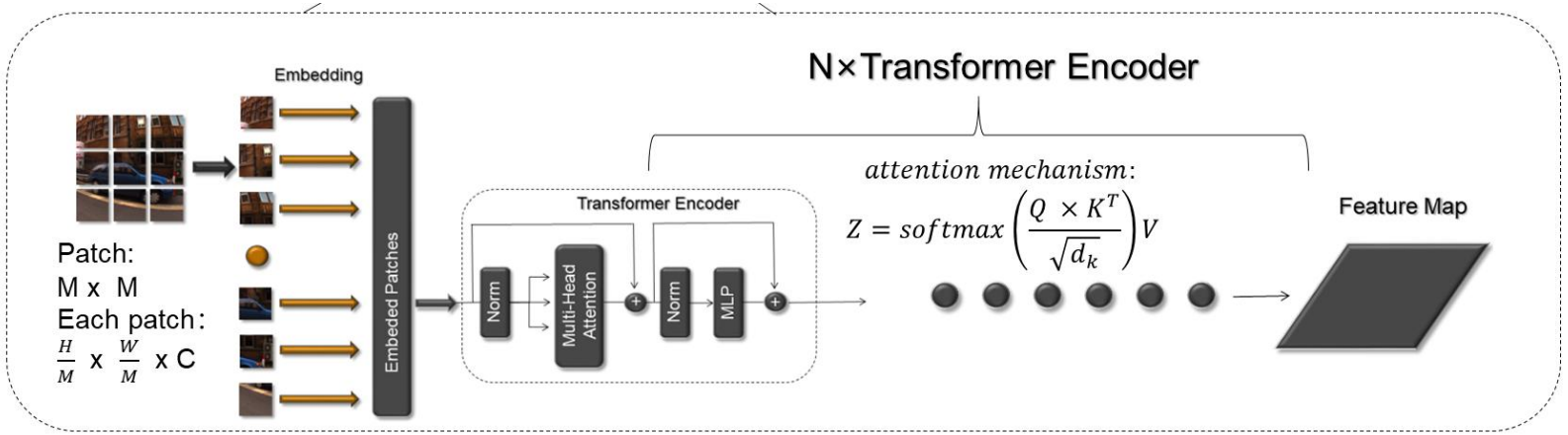


Positional Encoding



2 Transformer Based Loop Closure Detection (TLCD)

Transformer Encoder

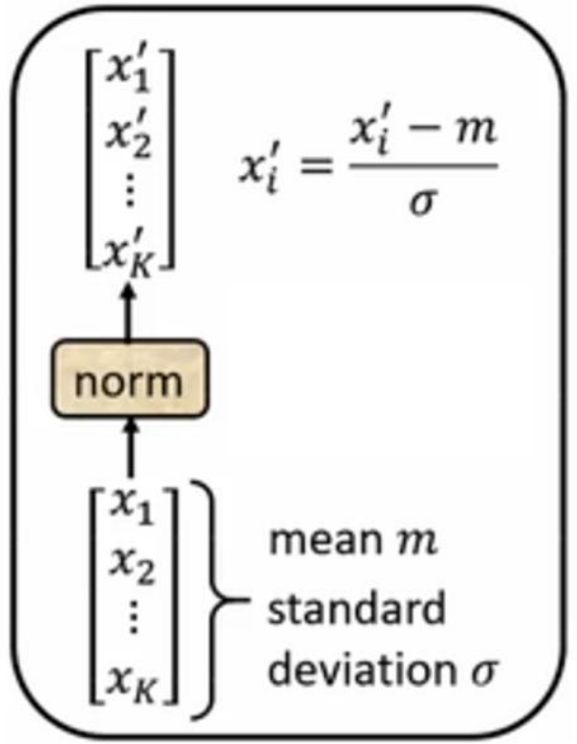


$$T_0 = [x_{class}, x_p^1 Emb, x_p^2 Emb, \ldots, x_p^N Emb] + Emb_{pos}$$

$$T_l' = LNorm(MHSA(T_{l-1}) + T_{l-1})$$

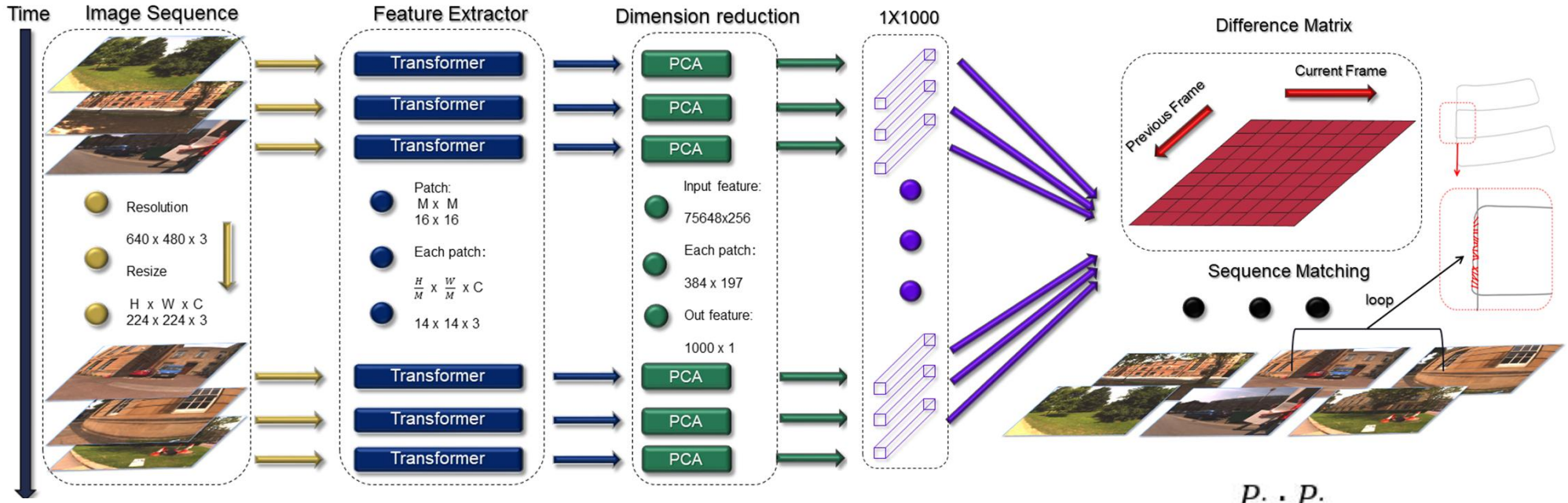
$$T_l = LNorm(MLP(T_l') + T_l')$$

Layer Normalization



2 Transformer Based Loop Closure Detection (TLCD)

Dimension Reduction & Sequence Matching



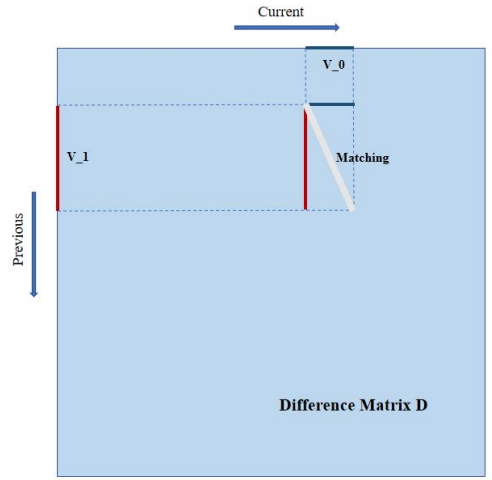
Output feature vectors of the transformer are too large to keep efficiency in the future feature matching process.

So we need to reduce the dimension of the feature without losing the information.

$$similarity = \cos(\theta) = \frac{P_i \cdot P_j}{\|P_i\| \|P_j\|}$$

Difference Matrix D: $D(i,j) = norm(1 - similarity)$

Setting different speed of potential sub-sequence for sequence total matching.



3 Experiments and Results

3 Experiments and Results

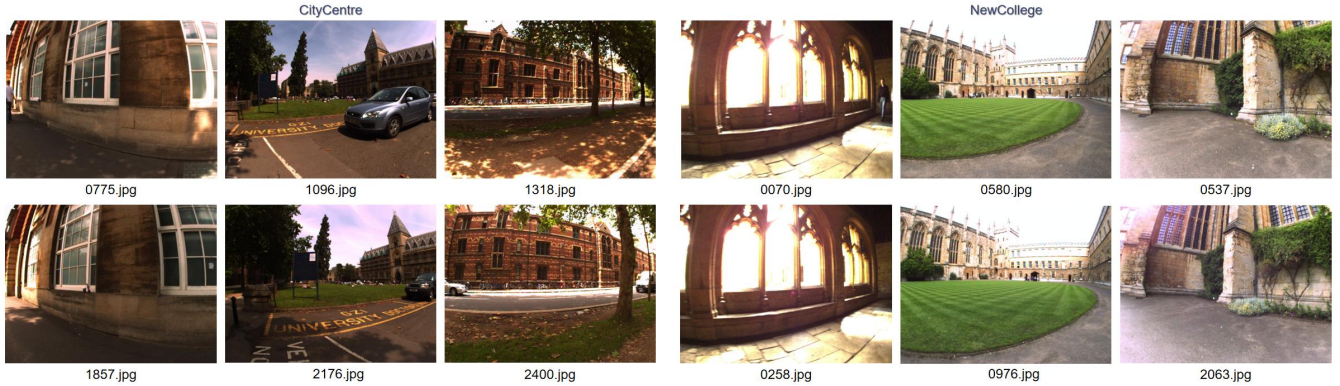
Datasets

Training Dataset Places365-standard

Large public dataset;
Scene classification;
365 categories;
1.8 milllion images.

LCD Datasets CityCentre and NewCollege

Small public datasets;
Robot carried two cameras in both side;
1.5 seconds per sample;
Contain real closed loop;
CityCentre, 2474 images;
NewCollege, 2146 images;
Ground truth matrix.



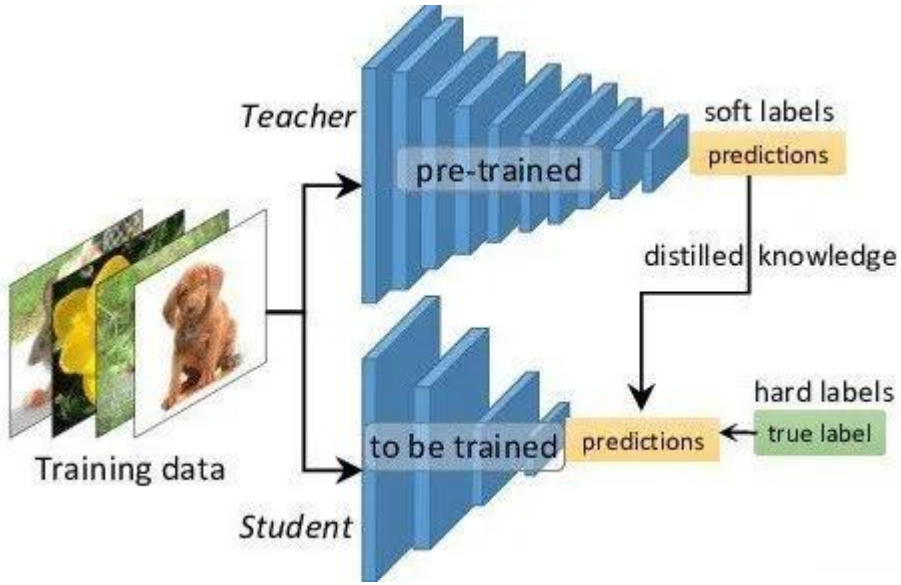
Knowledge Distillation in Transformer

No pre-trained model of transformer for scene classification.

Vit is hard to train, need large datasets.

Original Vit is too large to work on an edge device.

Knowledge distillation (Deit)



$$q_i = \frac{\exp\left(\frac{z_i}{T}\right)}{\sum_j \exp\left(\frac{z_j}{T}\right)}$$

For Deit, it uses convolutional neuarl network as the teacher model, and train the transformer student model, and finally get best performance.

In TLCD, we set well trained ResNet as the teacher model, and train the transformer.

As T approaches infinity, softmax output is more "soft". Therefore, a larger T can be used when training the student network. After training, normal T=1 was used for prediction. Minimize the cross entropy of the two distributions during training:

$$C = -p^T \log q$$

Training Result

Table 1 Training Results of Different Methods on Places365

Method	Top1	Top5
AlexNet	53.31%	82.75%
GoogLeNet	53.59%	84.01%
VGG	55.19%	85.01%
ResNet	54.65%	85.07%
Transformer	**53.28%**	**84.04%**

3 Experiments and Results

Precision-Recall Curve and Average Precision

Table 2 Classification of Loop Closure Detection

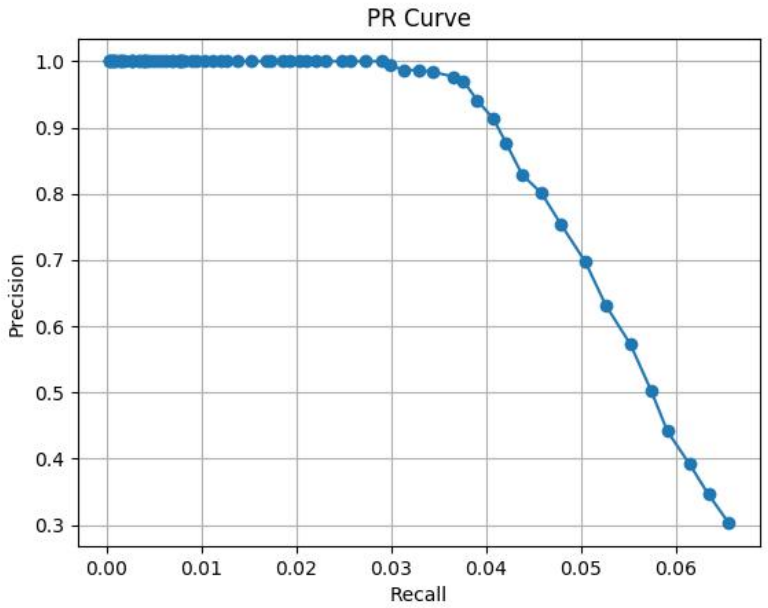
Prediction \ Groundtruth	Loop	Not Loop
Loop	**True Positive(TP)**	False Positive(FP)
Not Loop	False Negative(FN)	True Negative(TN)

False Positive(FP) Sample

False Negative(FN) Sample

$$Precision = \frac{TP}{TP + FP}$$

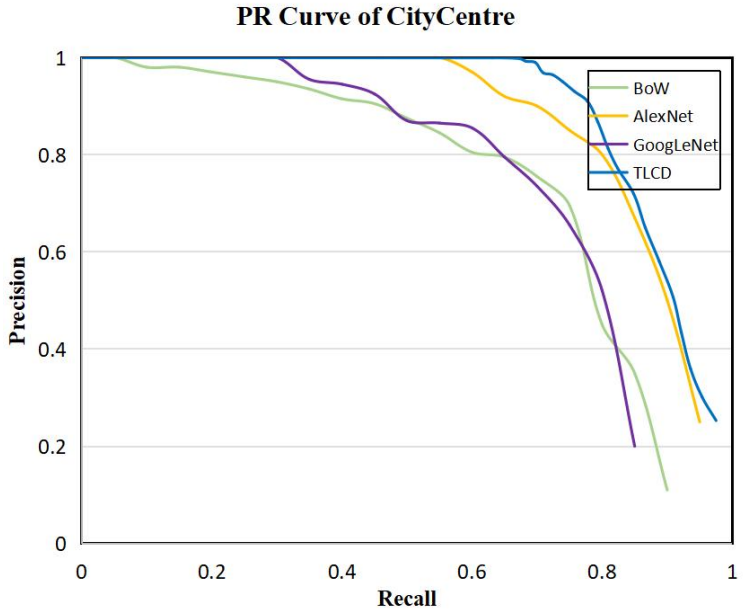
$$Recall = \frac{TP}{TP + FN}$$



Average Precision:

$$AP = \int_0^1 P(r)dr$$

3 Experiments and Results



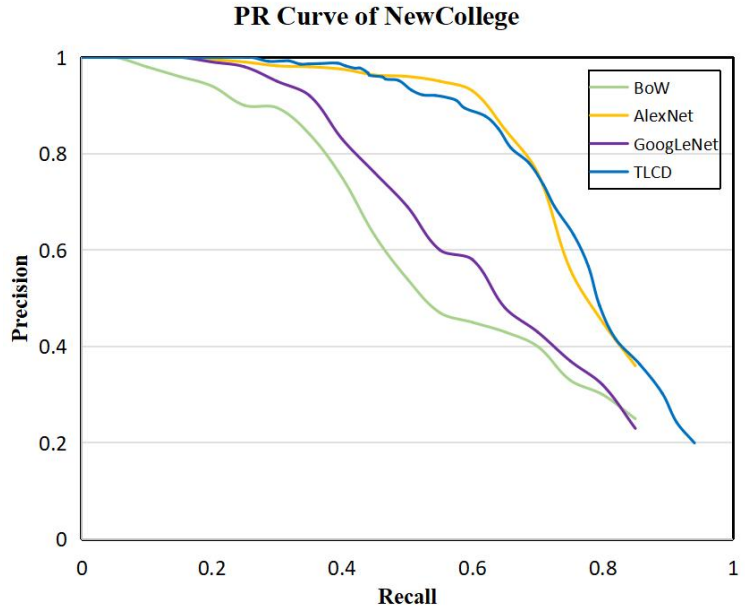
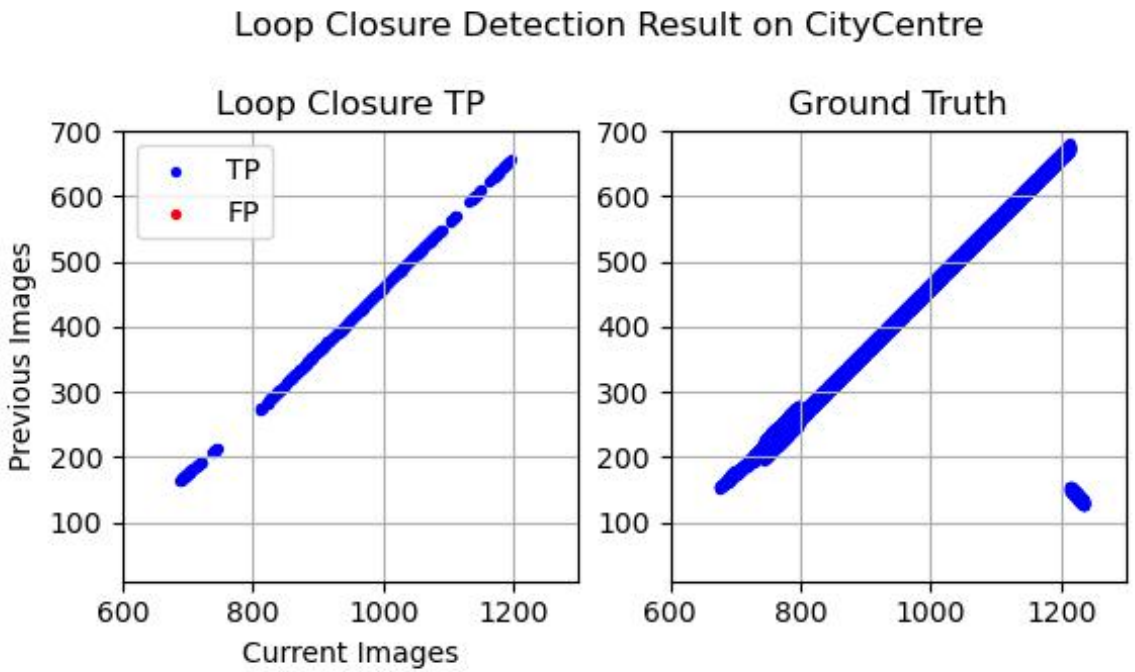
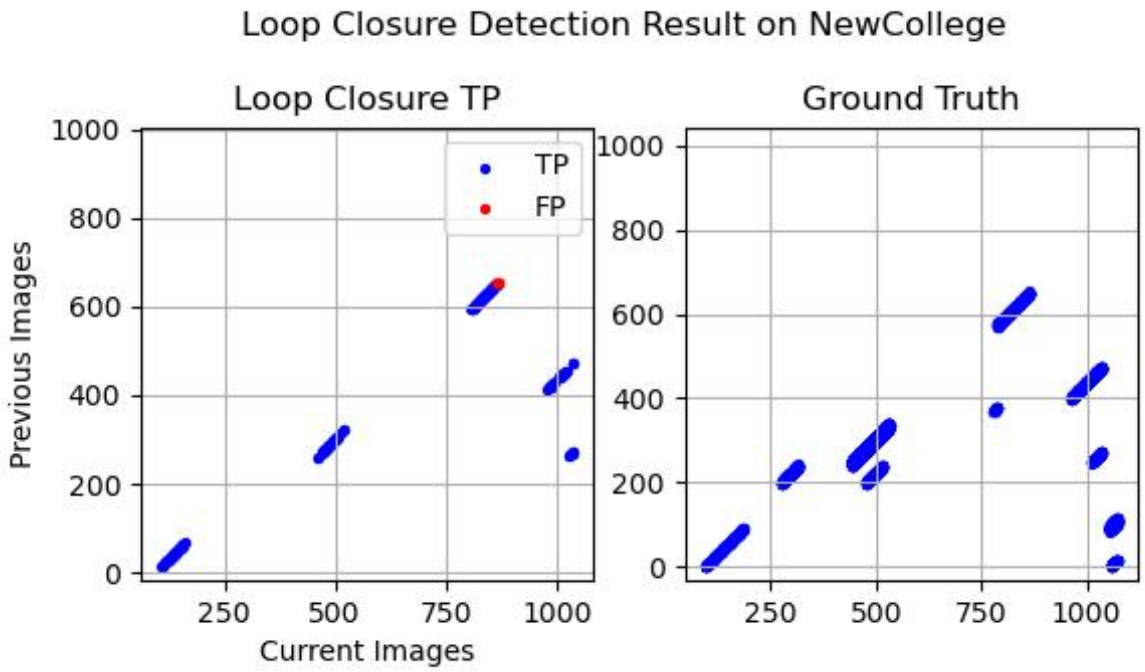


Table 3 Average Accuracy (AP) of TLCD Compared with Other Methods

Methods		Results	
		NewCollege	CityCentre
Traditional	BoW	62.38%	72.64%
	GIST	60.82%	69.79%
CNN-based	PCANet	73.76%	81.38%
	CaffeNet	74.04%	82.80%
	AlexNet	78.92%	85.87%
	GoogLeNet	66.47%	74.19%
	VGG	78.26%	82.13%
Transformer-based	Vision Transformer	**79.29%**	**89.05%**

3 Experiments and Results





Model Size and Time Consumption

Table 4 Comparation of the Model Size Between Different Model

Name	Structure	Size
AlexNet	CNN	>200MB
VGG	CNN	>500MB
ViT	Transformer	>300MB
TLCD	Distilled Transformer	**86MB**

Table 5 Time Consumption of TLCD When Evaluate on PC (CPU) (S/Frame)

Item		TLCD
CityCentre	Feature Extractor	0.948
	Total LCD	1.075
NewCollege	Feature Extractor	1.066
	Total LCD	1.186

4 Conclusion

4 Conclusion

In this paper, we propose a transformer-based loop closure detection algorithm (TLCD), which employs a distillation transformer as backbone to extract global features, and is combined with a sequence matching as back-end processing of principal component analysis (PCA) algorithm.

Results show that TLCD's average accuracy is up to **16.91%** higher than the traditional LCD method.	It is also about **3.18%** higher accuracy than the state-of-the-art convolutional neural network (CNN) based LCD method.


南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY
Thank you!