*Article*

# Self-supervised Foundation model based on Transformers for loop closure detection

**Mahmoud Saad** [1,†,‡], **Abdelmoniem bayomi** [2,‡]

[1] mahmoud.saadeldin101@gmail.com
[2] abayoumi@cu.edu.eg
* Correspondence: e-mail@e-mail.com; Tel.: (optional; include country code; if there are multiple corresponding authors, add author initials) +xx-xxxx-xxx-xxxx (F.L.)
† Current address: Affiliation 3.
‡ These authors contributed equally to this work.

**Abstract:** Loop closure detection is pivotal in enhancing Simultaneous Localization and Mapping (SLAM) for mobile robots by mitigating localization errors. Traditional techniques face challenges with environmental changes, driving the exploration of Visual SLAM (VSLAM) using Vision Deep Learning. This study explores SAM (Segment Anything Model), a self-supervised vision transformer, applied to loop closure detection. Leveraging SAM's encoded features and similarity comparison, our experiments on KITTI datasets demonstrate promising results. SAM shows adaptability to environmental changes, illuminating its potential as a foundational, plug-and-play model for robust loop closure detection.

**Keywords:** Foundation models. Segment Anything(SAM). Vision transformers. Self-supervised learning. Large Language Models (LLMs)

**Citation:** Lastname, F.; Lastname, F.; Lastname, F. Title. *Journal Not Specified* **2024**, *1*, 0. https://doi.org/

Received:
Revised:
Accepted:
Published:

## 1. Introduction

Simultaneous Localization and Mapping (SLAM) is a transformative field in robotics and autonomous navigation, revolutionizing environmental perception and navigation for machines. This multidimensional process of concurrently determining a robot's position while constructing a map of its surroundings has propelled advancements across various domains, from enhancing self-driving vehicles to enabling the autonomy of aerial drones. However, at the core of SLAM lies an intricate and indispensable challenge—loop closure detection.

Loop closure detection acts as the linchpin in fortifying the integrity of mapped environments by addressing the tendency of mobile robots to accumulate localization errors over time, commonly referred to as Localization Drift[1]. This challenge arises from the compounding effect of even minor errors in a robot's localization estimation, leading to substantial deviations in its perceived position over prolonged operations. The pivotal role of loop closure detection lies in recognizing and reconciling when a robot revisits a previously traversed location, thereby enabling the correction of accumulated localization errors.

Traditional methodologies for loop closure detection often confront formidable obstacles in coping with dynamic environmental changes, such as abrupt illumination alterations or the disappearance and reappearance of objects within the robot's field of view. These complexities render singular techniques insufficient in addressing the intricate nature of environmental variability. As a response to these challenges, a burgeoning body of research has converged towards the integration of Visual SLAM (VSLAM), harnessing the potential of Vision Deep Learning techniques.

The current existing existing Visual deep learning techniques are trying to make powerful architecture to build the most useful features from input frame. But limitation for that. is generalization for being plug-and-play in any environment. for better generalization of any environment, It's required to be training with huge amount of data with attention model architecture. And that could be founded in attention-based transformers. and the huge data used for training SAM.

In this realm of evolving methodologies, this paper introduces a pioneering approach—the "Segment Anything Model" (SAM)[2]—crafted explicitly to tackle the exigencies of loop closure detection within the domain of Visual SLAM. Leveraging the advancements in Foundation Large models witnessed in domains like natural language processing (NLP), our methodology employs vision-based Foundational models, exemplified by SAM, the "Segment Anything Model". SAM stands as a testament to the transformative power of self-supervised learning and Vision Transformers (ViT), embodying insights gleaned from a substantial corpus of training data, encompassing over 11 million images and an astounding one billion masks.

SAM, structured with both encoder and decoder architecture, epitomizes innovation by focusing solely on extracting image embeddings through its encoder. This approach hinges on measuring the similarity between these image embeddings, revolutionizing the landscape of loop closure detection in SLAM. As we navigate through the intricacies of SAM, this paper endeavors to provide a comprehensive understanding of its architectural nuances, strategic deployment, and performance evaluation. Additionally, we contextualize SAM within the broader scope of loop closure detection solutions by elucidating related works. Finally, we present empirical findings that shed light on SAM's robust capabilities, inherent limitations, and avenues for future research.

## 2. Related Work

Loop closure detection has been a central concern in the field of Simultaneous Localization and Mapping (SLAM) for many years. Its significance in preventing cumulative localization errors and maintaining map consistency has driven extensive research efforts. This section provides an overview of the key methodologies and approaches explored in the pursuit of robust loop closure detection.

### *2.1. Traditional Methods*

Traditional methods for loop closure detection encompass a wide spectrum of techniques. Among the well-known strategies are scan matching[3], feature-based methods[4], and bag-of-words[5] approaches. Scan matching techniques, such as Iterative Closest Point (ICP), have been instrumental in comparing LIDAR scans to identify potential loop closures. Feature-based methods involve extracting distinctive features from sensor data and matching them across different time steps. Meanwhile, bag-of-words techniques exploit visual words to represent image content and assess the similarity between images. These traditional methods have demonstrated competence in various scenarios and settings.

### *2.2. Graph-Based Approaches*

Graph-based SLAM approaches[6], particularly in the form of pose graph or factor graph optimization, have been widely employed for loop closure detection. In these methods, loop closures are modelled as constraints within a graph structure. Optimization techniques, such as Bundle Adjustment, are applied to refine the robot's trajectory and the estimated map, accounting for loop closures. The success of graph-based methods in addressing loop closure issues is well-documented in the literature.

### *2.3. loop Closure detection with CNN*

In recent years, integrating deep learning techniques has ushered in a new era of loop closure detection in SLAM. Convolutional Neural Networks (CNNs), recurrent networks, and more recently, Vision Transformers (ViT), have been applied to this task. These methods learn discriminative features from sensor data, facilitating loop closure detection even in challenging environments. Notably, the advancement of self-supervised learning has enabled models to extract meaningful representations from vast amounts of unlabeled data.

PlaceNet [7], emerges as an innovative, adaptable model catering to visual loop closure detection. This multi-scale deep autoencoder network is enriched by a semantic fusion layer, enhancing scene comprehension. The core concept of PlaceNet revolves around discerning critical areas within dynamic scenes, steering clear of distractions posed by moving elements to prioritize scene landmarks. Training PlaceNet involves discerning dynamic elements through the acquisition of grayscale semantic maps, pinpointing static and moving objects in images. The outcome is the generation of robust, scale-invariant, semantic-aware deep features well-suited for dynamic environments.

LoopNet [8], a novel plug-and-play model, LoopNet, to find similarities between scenes via determining key landmarks to focus on without being distracted by scene variations. throught multi-scale attention-based Siamese convolutional model learns feature embeddings that focus on the important objects in the scene that can help to be robust to illumination change instead of general features.

MATC-Net [9],a multi-scale asymmetric temporal convolution network (MATC-Net), which generates sequential features and transformed global features through its aggregation branch and transformation branch, respectively. with those extracted features which can be fused, a MATC-Net-based hierarchical LCD framework including two similarity measurement processes is constructed.

### *2.4. Loop closure detection using Transformer*

While Vision Transformers have demonstrated remarkable success in various computer vision tasks, they have also found their place in loop closure detection. Several approaches have utilized ViT to extract image embeddings and perform similarity measurements for loop closure identification. The evolution of Vision Transformer-based solutions has sparked substantial interest in exploring their capabilities in SLAM contexts.

TLCD [8], proposes a transformer-based loop closure detection algorithm (TLCD), which utilizes a distillation transformer as its core for global feature extraction, TLCD integrates sequence matching as the backend process using the principal component analysis (PCA) algorithm. TLCD demonstrates an adept ability to generate Precision-Recall curves based on various public datasets, including CityCentre and New-College datasets. Findings reveal TLCD's superior performance, showcasing an average accuracy surpassing the traditional LCD method by up to 16.91%. Moreover, it outperforms the state-of-the-art CNN-based LCD method by approximately 3.18% in accuracy.

HiTPR [10], Hierarchical Transformer for Place Recognition in Point Cloud. This study presents a hierarchical transformer-oriented strategy crafted to tackle the intricacies involved in recognizing locations within point cloud data. Utilizing transformers, the authors harness their capability to comprehend intricate associations within the point cloud, facilitating precise identification of previously accessed sites. HiTPR's pioneering approach in place recognition holds promise for enhancing loop closure detection in Simultaneous Localization and Mapping (SLAM), presenting a valuable avenue for exploring sophisticated methodologies in spatial comprehension based on point cloud data.

## 3. Our Approach

A Foundational large model -popular in the Natural language processing field as LLMs -is a type of Deep learning model notable for its ability to achieve general-purpose understanding and generation. Foundational models acquire these abilities by using massive amounts of data to learn billions of parameters during training and consuming large computational resources during their training and operation. Most of the time they are based on Self-supervised learning and transformer architecture. As huge success happens in the AI field due to those foundation models, We proposed a known foundational model called SAM to be a general solution for LCD.

### *3.1. Segment Anything model SAM*

Powerful zero-shot learning and few-show learning foundation model, SAM. Follows the encoder-decoder architecture. It consists of three main modules: image encoder, prompt encoder and mask decoder. Given an input image, the image encoder outs an image embedding. SAM implements and adapts a pre-trained ViT-H/16 masked auto-encoder. This is a relatively large model with strong performance. Before passing the image embeddings to the Mask decoder to output a valid segmentation mask. A set of prompt embeddings is generated from the Prompt encoder to give the mask decoder more context about what regions of the images need to be focused on. The prompt encoder takes sparse prompts (i.e. points, boxes and text) that are translated into embedding vectors. It also accepts to take a mask referring to a region of interest in the image. Those masks are simply downsampled with stride convolutions and added with the image embeddings.

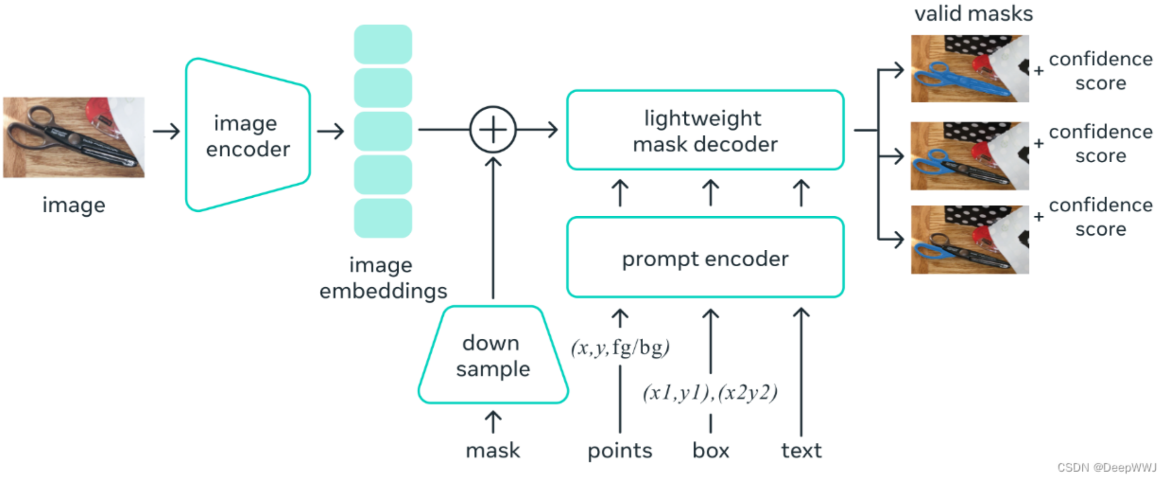

**Figure 1.** SAM overview architecture.

### *3.2. Similarities of Encoded Features and Decoder Prompting*

As discussed, SAM can generate rich semantic labels for any given image. Due its powerful capabilities, SAM is promising to work in any environment. Given the advantage of SAM model. The Remaining work builds full Loop closure detection is to find a generic similarity technique to inform if the scene is similar to the seen scene. If two frames of images are similar, they must contain numerous identical semantic labels. Therefore, keyframes with large differences can be roughly excluded through similarity comparison of semantic labels.

In this experiment, We only conducted the similarity on the feature vector extracted. simply, Using the embedding extracted from SAM encoder, Cosine similarity is used to select the candidate frames from the seen scene. And Semantic labels similarity are left as future work

### *3.3. Overall System Flow*

first, Segment anything take the image as input and process it throught it's feature extraction enconder layer, and return feature verctor represnting the frame. This vector is very descriptive to image and it's content. As the encoder is so powerful. We just passed this vector to cosine similarity measure with history vectors to be able get Loop closure indicator. . In the long trip there will be an issue storing all history frames. this will be intensive in memory usage for that, We experimented with KITTI dataset. for sequencial frames they will be very similar logically. for that, We took sample of a frame and drop X next frames due the camera capture rate.

| Dataset | Approach | Precision % | Recall % |
|---|---|---|---|
| Sequence 00 | Kim [11] | 100 | 87 |
| Sequence 00 | Gálvez-López[12] | 100 | 92 |
| Sequence 00 | Proposed | NA | NA |
| Sequence 02 | Kim | 90 | 73 |
| Sequence 02 | Gálvez-López | 100 | 80.6 |
| Sequence 02 | Proposed | NA | NA |
| Sequence 05 | Kim | 100 | 90 |
| Sequence 05 | Gálvez-López | 100 | 87.6 |
| Sequence 05 | Proposed | 100 | 40 |

**Table 1.** Comparison with existing methods

## 4. Experimental Results

In this section we provide the quantitative analysis about usage of SAM without any training. We experiment with KITTI dataset for loop closure. KITTI dataset serve many research in autonomous robotics as it came up with various perception and labelled records. In KITTI dataset, the odametry data provide 11 sequences for camera motion. seven of them contains loop closure. SAM is tested with sequences of 00, 02 and 05. sequence 02 is a challenging sequence as it has forward and reverse visit. Meanwhile, the other selected sequence only have forward visit.

The performance of SAM is evaluated using precision and recall values. precision tell how much the model is accurate when it tell there is a loop closure detected. while the recall reports the how much the model detect loop closure compared to the actual number of loop closure.

for sequence 05, the Model is tested with two threshold, first with 0.9 and with 0.85. Each show different precision recall values that helped us to reach a conclusion. for threshold equal 0.9 we achieved precision of 100% while we got 10% in recall value on the other hand, when applying threshold equal to 0-85, the model achieved precision of 75% and recall value 99.2%. compared to baseline models, the other models might be achieved better balance between precision and recall. But overall, the proposed approach might show better generalization for any scene can be applied to.

### *4.1. result analysis*

From the result we got, we can say SAM model is showing great potential to be generalized approach for Loop closure. the model is trained over million of data to be able to segment correctly any given image. If we focused on model object identify space. we can see the model is Learned to separate object's like cars seas, ships, houses. and make similarity objects of the same type near to each other of the objects of the same type near to each other. Given 2 images contains objects of same type the similarity of encoder feature vector should be near.mean while comparing with another image contains different object, the similarity of encoder feather should be low. from that, we can say the given images in the records to the model is contains always street, sky, trees, signs. which all can be very near in the SAM feature space. As they contain similar overall types. despite that, In threshold of 0.85 we can see the model is able to precise tell if loop closure happened with 75%. with almost not loosing any loop behind. and very accurate with threshold of 0.9.

## 5. Conclusions

Loop closure detection is a pivotal task in SLAM, ensuring the consistent and accurate construction of environmental maps. Our proposed method, the "Segment Anything Model" (SAM), capitalizes on the power of self-supervised learning and Vision Transformers (ViT). SAM boasts a profound knowledge base. While SAM's architecture includes both an encoder and decoder, we focus on leveraging the encoder to extract image embeddings. The crux of our approach centers on employing similarity measurements to identify loop closures, offering a fresh perspective on a long-standing challenge. Through a series of rigorous experiments, we demonstrate the efficacy and generalization potential of SAM in loop closure detection within the domain of SLAM.

## 6. Future work

There are many rooms for improvement and experimenting the power and capabilities of SAM. Using the prompting encoder and decoder to inform more context about the object in the frame. Semantic labels similarity are left as future work and it might be promising as many works used segmentation masks for filtering out candidates frames

## 7. References

[1] A. Angeli, D. Filliat, S. Doncieux, and J.-A. Meyer, “A fast and incremental method for loop-closure detection using bags of visual words,” IEEE Transactions on Robotics, pp. 1027–1037, 2008.

[2] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, “Segment Anything,” arXiv:2304.02643, 2023.

[3] J. Li, H. Zhan, B. M. Chen, I. Reid and G. H. Lee, "Deep learning for 2D scan matching and loop closure," 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 2017, pp. 763-768, doi: 10.1109/IROS.2017.8202236.

[4] Azzam, R., Taha, T., Huang, S. et al. Feature-based visual simultaneous localization and mapping: a survey. SN Appl. Sci. 2, 224 (2020). https://doi.org/10.1007/s42452-020-2001-3

[5] Z. Huishen, X. Ling, Y. Huan and W. Liujun, "An improved bag of words method for appearance based visual loop closure detection," 2018 Chinese Control And Decision Conference (CCDC), Shenyang, China, 2018, pp. 5682-5687, doi: 10.1109/CCDC.2018.8408123.

[6] Duan, Ran, Yurong Feng, and Chih-Yung Wen. 2022. "Deep Pose Graph-Matching-Based Loop Closure Detection for Semantic Visual SLAM" Sustainability 14, no. 19: 11864. https://doi.org/10.3390/su141911864

[7] H. Osman, N. Darwish, and A. Bayoumi, “PlaceNet: A multi-scale semantic-aware model for visual loop closure detection,” Engineering Applications of Artificial Intelligence, vol. 119, p. 105797, 2023.

[8] H. Osman, N. Darwish, and A. Bayoumi, “LoopNet: Where to focus? Detecting loop closures in dynamic scenes,” IEEE Robotics and Automation Letters, vol. 7, no. 2, pp. 2031–2038, 2022.

[9] F. Fu, J. Yang, J. Zhang, and J. Ma, “MATC-Net: Learning compact sequence representation for hierarchical loop closure detection,” Engineering Applications of Artificial Intelligence, vol. 125, p. 106734, 2023.

[10] Z. Hou, Y. Yan, C. Xu, and H. Kong, “HiTPR: Hierarchical transformer for place recognition in point cloud,” in 2022 International Conference on Robotics and Automation (ICRA), 2022, pp. 2612–2618.

[11] G. Kim and A. Kim, “Scan context: Egocentric spatial descriptor for place recognition within 3d point cloud map,” in 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018, pp. 4802–4809.

[12] C. Li, H. Ren, M. Bi, C. Ding, W. Li, R. Zhang, X. Liu, and H. Yu, “TLCD: A Transformer based Loop Closure Detection for Robotic Visual SLAM,” in 2022 International Conference on Advanced Robotics and Mechatronics (ICARM), 2022, pp. 261–267.