MUST-Loc: Multi-view Uncertainty-aware Semantic Token Association for Object-level Global Localization

Gihyeon Lee[1], Young-Sik Shin[2†], and Younggun Cho[1†]

***Abstract*— Object-level global localization is highly sensitive to semantic uncertainty from viewpoint variations in open-set scenarios. To address this problem, we present MUST-Loc, a multi-view, uncertainty-aware semantic token matching framework. The key idea is to aggregate object-level tokens through online updates in the mapping process to form mean–variance descriptors, capturing viewpoint-induced variability while maintaining semantic consistency. At the localization query, we compute uncertainty-aware semantic similarity, which down-weights high-variance token dimensions to establish reliable correspondences under semantic ambiguity. Finally, the camera pose is estimated by selecting the solution that maximizes the Wasserstein-based alignment score between observed detections and projected landmark hypotheses. For rigorous validation, we evaluate on challenging TUM RGB-D sequences with occlusions, label noise, and diverse categories, showing consistent improvements over baselines in association and pose accuracy. Project page: https://leekh951.github.io/MUST-Loc.**

I. Introduction

In open-world navigation, robots must operate reliably despite unknown objects, complex environments, and ambiguous observations. Recent advances have improved the robustness of navigation systems. However, semantic ambiguity and uncertainty inherent in human-centric environments remain significant challenges. In particular, object-level global localization, a key component for robust autonomy, is highly sensitive to semantic uncertainty caused by viewpoint variations in open-set scenarios [1].

To address this, recent methods [2, 3] have moved beyond predicted labels by integrating embeddings from vision and language model (VLM), which can recognize unknown or previously undefined object classes. Nevertheless, similar to graph-based methods that rely on semantic labels [4–8], these approaches typically assign only a textual caption to each landmark. This design fails to capture semantic variations across multiple viewpoints. Therefore, there is a critical need for object-matching approaches that aggregate semantic features from multiple views into a probabilistically interpretable descriptor, adaptable to arbitrary query observations.

In this paper, we propose *MUST-Loc*, a multi-view uncertainty-aware semantic token matching method for global localization. Our key idea is to incrementally compute the mean and variance of semantic token vectors for object landmarks during mapping and to leverage these statistical descriptors to resolve semantic ambiguity during query matching, as illustrated in Fig. 1.

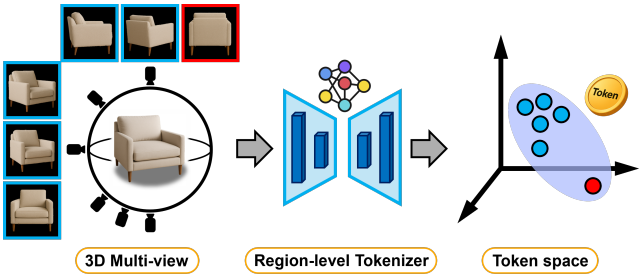


Fig. 1. **Multi-view inconsistency.** Object observations from different viewpoints often produce inconsistent semantic tokens, leading to scattered embeddings in the token space.

Specifically, we design a similarity measure between landmarks and query observations, where semantic variance tokens act as gating weights to modulate distributional similarity. This enables more accurate and robust object correspondence for global localization.

Our main contributions are summarized as follows:

- **Semantic Disambiguation:** We show that aggregating semantic tokens across multiple views mitigates ambiguity caused by viewpoint variations.
- **Uncertainty-aware Matching:** We introduce a similarity measure that incorporates variance tokens to provide probabilistic weighting in distributional comparisons.
- **Robust Global Localization:** On public dataset [9] with clutter, occlusion, and label noise, our method yields consistent gains in association and pose, enabling robust global localization.

II. Related Work

A. Data Association for Global Localization

Object-level data association for global localization largely falls into two categories: label-based and VLM-based methods. Label-based approaches rely on predicted semantic labels from pretrained detectors, but single-label dependence often introduces ambiguity. To mitigate this, graph-based extensions [4–8] incorporate contextual labels and topological relations, which improve robustness under viewpoint changes. However, these methods remain tied to graph structures and overlook object-inherent uncertainty. VLM-based methods instead leverage embedding features to capture more generalized vision–language relationships. For instance, Matsuzaki et al. [2] annotate map landmarks with natural language and match them via conceptual similarity, while

[1]Gihyeon Lee, and [1†]Younggun Cho are with the Electrical and Computer Engineering, Inha University, Incheon, South Korea leekh951@inha.edu, yg.cho@inha.ac.kr

[2†]Young-sik Shin is with Depart. of AI Machinery, Korea Institute of Machinery and Materials, Daejeon, South Korea yshin86@kimm.re.kr

Corresponding Authors: Young-Sik Shin and Younggun Cho

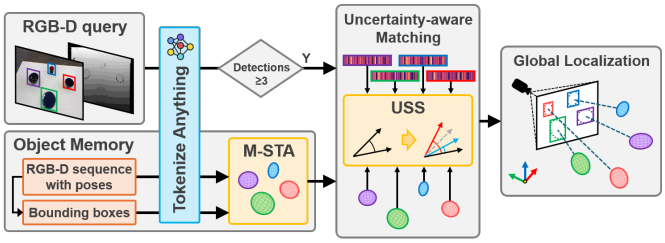


Fig. 2. **Overview of our method.** Our method aggregates multi-view semantic tokens from TAP [16], matches them via uncertainty-aware similarity, and estimates the camera pose based on a Wasserstein alignment score.

Matsuzaki et al. [3] combine CLIP [10] embeddings with Semantic histogram [11] and perform clique-based graph matching. Although effective, these approaches still struggle in scenes with semantically similar objects and insufficiently account for multi-view uncertainty. In contrast, our method probabilistically models multi-view uncertainty of semantic tokens directly aggregated from objects, achieving robust and discriminative association without explicit graph structures.

B. Vision-Language Models

Transformer [12]-based foundation models have advanced vision-language reasoning by pretraining on large-scale image text pairs. Among them, CLIP [10] enables open-vocabulary scene understanding and has been applied to navigation [13] and place recognition [14, 15]. Nevertheless, its embeddings derived from cropped bounding boxes often include background, reducing discriminability and consistency across viewpoints. Tokenize Anything via Prompting (TAP) [16], by contrast, extracts embeddings directly from object masks, suppressing background noise and providing more consistent, uncertainty-aware features across views. Leveraging these strengths, we adopt TAP to enhance semantic robustness and discriminability in object-level global localization.

III. MUST-LOC

A. Overview

As illustrated in Fig. 2, the method comprises multi-view semantic token aggregation, uncertainty-aware data association, and pose estimation. For each landmark o_j, tokens from sequential views are aggregated to update mean–variance descriptors $(\boldsymbol{\mu}_j, \boldsymbol{\sigma}_j^2)$, capturing viewpoint-dependent variability while preserving semantic consistency. Given detections d_i, Uncertainty-aware Semantic Similarity (USS) is computed between each d_i and candidate landmarks o_j, selecting correspondences $(d_i, o_j) \in \mathcal{P}$ robust to semantic ambiguity. Finally, the camera pose T is estimated from $\mathcal{P}$ by maximizing a Wasserstein alignment score between observed detections and projected landmark hypotheses.

B. Uncertainty-aware Semantic Descriptor

1) TAP-based Tokenization: For each object, we employ the semantic token from TAP [16]. This token encapsulates region-level visual features and a semantically aligned distribution:

$$\mathbf{v}_i = \mathrm{TAP}(I, B_i), \tag{1}$$

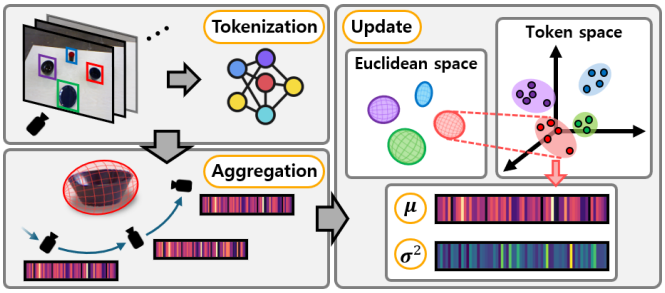


Fig. 3. **Multi-view Semantic Token Aggregation (M-STA).** Our method incrementally updates mean and variance tokens to obtain semantically consistent and uncertainty-aware object descriptors.

where $\mathrm{TAP}(\cdot,\cdot)$ denotes a promptable region-level tokenizer that segments the region specified by B_i and produces a semantic token. I and B_i denote the query RGB image and the bounding box, respectively. As TAP prompts, the bounding boxes are inferred under open-set conditions using GroundingDINO [17].

2) Multi-view Semantic Token Aggregation: Human-made objects exhibit large variations in observed shape across viewpoints, yet their core semantic attributes remain invariant. For example, a cup consistently retains its fundamental property as a container for liquids regardless of the viewing angle. However, a robot observing an object from a single viewpoint often fails to capture its complete semantic characteristics, motivating the integration of multi-view evidence into a unified descriptor.

As illustrated in Fig. 3, we introduce Multi-view Semantic Token Aggregation (M-STA). In the mapping process, we incrementally update the mean and variance of object tokens using Welford's online algorithm:

$$\boldsymbol{\mu}_{j,t} = \boldsymbol{\mu}_{j,t-1} + \frac{\mathbf{v}_{j,t} - \boldsymbol{\mu}_{j,t-1}}{t}, \tag{2}$$

$$\boldsymbol{M}_{j,t} = \boldsymbol{M}_{j,t-1} + (\mathbf{v}_{j,t} - \boldsymbol{\mu}_{j,t-1}) \odot (\mathbf{v}_{j,t} - \boldsymbol{\mu}_{j,t}), \tag{3}$$

$$\boldsymbol{\sigma}_{j,t}^2 = \frac{\boldsymbol{M}_{j,t}}{t-1}, \tag{4}$$

where $\boldsymbol{\mu}_{j,t}$ and $\boldsymbol{\sigma}_{j,t}^2$ denote the running mean and variance vector after t observations, $\boldsymbol{M}_{j,t}$ is the accumulated sum of squared deviations used to compute the variance, and $\odot$ is the Hadamard product.

The mean token encodes the consensus of semantic information, while the variance token quantifies uncertainty across views. By explicitly modeling both, M-STA generates object descriptors that are semantically robust and account for uncertainty across views.

C. Uncertainty-aware Semantic Similarity

Semantic similarity between query observations and map landmarks can be severely distorted when feature dimensions exhibit high uncertainty. To address this, we design a similarity measure, referred to as USS, that explicitly incorporates variance information. Building on the multi-view aggregated mean $\boldsymbol{\mu}_j$ and variance $\boldsymbol{\sigma}_j^2$ tokens, we define an uncertainty-aware similarity measure S with the query object token $\mathbf{v}_i$

as follows:

$$S(\mathbf{v}_i, \boldsymbol{\mu}_j) = \frac{\mathbf{v}_i^\top \Lambda_j \boldsymbol{\mu}_j}{\sqrt{\mathbf{v}_i^\top \Lambda_j \mathbf{v}_i}\,\sqrt{\boldsymbol{\mu}_j^\top \Lambda_j \boldsymbol{\mu}_j}}, \tag{5}$$

where the diagonal weight matrix Λ_j is defined from the element-wise uncertainty of the mean token vector $\boldsymbol{\mu}_j$:

$$\Lambda_j = \mathrm{diag}(e^{-\lambda \boldsymbol{\sigma}_j^2}), \tag{6}$$

with $\mathrm{diag}(\cdot)$ constructing a diagonal matrix from a vector argument, and λ is a scaling hyperparameter that serves to exponentially down-weight dimensions with higher variance.

This formulation explicitly integrates uncertainty into the similarity computation, enabling robustness against noisy or ambiguous features. Based on this robust similarity, we select the top-1 landmark o_j for each query detection d_i to construct the set of correspondences $\mathcal{P}$.

D. Global Localization

A stochastic iterative approach is executed for N iterations. At each iteration, we randomly sample a previously unselected set of three unique correspondences, $\mathcal{P}_{samp}$, from $\mathcal{P}$. Given the $\mathcal{P}_{samp}$ that satisfies the required conditions, a candidate camera pose $\tilde{T}$ is estimated by solving the Perspective-3-Point (P3P) problem. For each $(d_i, o_j) \in \mathcal{P}$, the dual quadric of o_j is transformed by $\tilde{T}$ to generate a projected prior bounding box B_j. Both query bounding box B_i and the corresponding B_j are modeled as Gaussian distributions, denoted as $D_i = G(\mu_i, \Sigma_i)$ and $D_j(\tilde{T}) = G(\mu_j, \Sigma_j)$, respectively. The degree of alignment between the D_i and D_j is measured by the normalized Wasserstein distance proposed by [18] as follows:

$$W_n\big(D_i, D_j(\tilde{T})\big) = \exp\left(-\frac{\sqrt{W_2^2\big(D_i, D_j(\tilde{T})\big)}}{C}\right), \tag{7}$$

where $W_2^2(D_i, D_j)$ is the 2^{nd} order Wasserstein distance between two Gaussians, and C is a scale factor. This iterative process produces a set of N candidate poses, denoted as $\mathcal{T} = \{\tilde{T}^{(k)}\}_{k=1}^{N}$. Finally, the camera pose T is estimated by maximizing the Wasserstein alignment score as follows:

$$T = \underset{\tilde{T} \in \mathcal{T}}{\arg\max}\, \frac{1}{|\mathcal{P}|} \sum_{(d_i, o_j) \in \mathcal{P}} W_n\big(D_i, D_j(\tilde{T})\big). \tag{8}$$

IV. Experimental Results

A. Setup

To evaluate the proposed methodology, we adopt the TUM RGB-D public benchmarks [9]. Specifically, we employ the 'Fr2_dishes' sequence to assess discriminative capability among visually similar objects. Furthermore, to examine generalization under complex object arrangements, we utilize the 'Fr2_desk' and 'Fr2_person' sequences. Notably, 'Fr2_person' shares the same underlying object layout as 'Fr2_desk' but incorporates dynamic factors such as human presence and object relocation, thereby providing a more challenging evaluation scenario.

TABLE I

Data Association and Pose Estimation Performance

Dataset	Method	Data Association		SR_{succ} [%]↑			TE [m]↓
		F1↑	MOTA↑	@0.5m	@1m	@2m	
Fr2_dishes	GOReloc	0.993	0.758	59.96	64.81	90.83	0.6847
	Ours	**0.998**	**0.996**	**98.99**	**99.07**	**99.40**	**0.0865**
Fr2_desk	GOReloc	0.866	0.694	52.34	68.49	78.99	1.0436
	Ours	**0.888**	**0.769**	**96.37**	**97.11**	**99.17**	**0.1774**
Fr2_person	GOReloc	0.738	0.428	27.71	41.11	53.57	1.9972
	Ours	**0.933**	**0.853**	**82.01**	**87.82**	**90.40**	**0.5762**

TABLE II

Scene Invariance Analysis in Cross-session Scenario

Dataset	Method	SR_{all} [%]↑				SF↑
		@0.5m	@1m	@2m	@5m	
Fr2_person	ORB-SLAM2	8.68	8.68	8.68	8.68	353
	GOReloc	21.76	32.28	42.07	75.14	3194
	Ours	**72.66**	**77.80**	**80.08**	**87.63**	**3603**

We evaluate data association using the F1 score and MOTA [19]. For pose estimation, we report the mean translation error (TE) alongside two success-rate metrics: SR_{succ}, the proportion of success frames (SF) with TE below thresholds of $\{0.5, 1, 2, 5\}$ m, and SR_{all}, the success rate over the entire sequence.

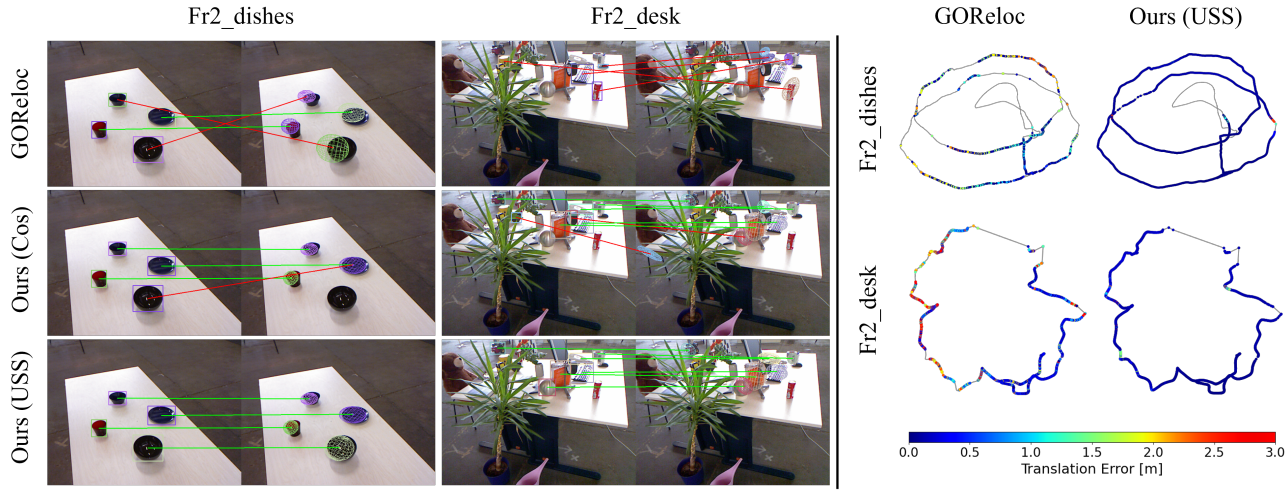
As primary baselines, we adopt the semantic graph–based method GOReloc [8] and, to assess scene invariance, ORB-SLAM2 using feature points [20].

B. Data association and Pose Estimation

Table I presents the performance comparison between baseline and the proposed approach on multi-object data association and pose estimation under ambiguous, complex scenarios with substantial scene variations. Table II reports scene invariance analysis on cross-session sequences with dynamic elements.

1) Evaluation under Object Ambiguity Scenarios: In the 'Fr2_dishes' sequence of Table I, the baseline exhibits low MOTA, which we attribute to semantic ambiguities among objects. In particular, while our method shows consistently high SR_{succ} values with variations within 1%, the baseline shows lower robustness under stricter error thresholds. The results suggest that our method more effectively accounts for semantic uncertainty, whereas the baseline relies on graph structures constructed from closed-set labels.

2) Evaluation under Object Complexity Scenarios: In the 'Fr2_desk' and 'Fr2_person' sequences, which are dominated by severe occlusions and partial detections, our approach consistently achieved improvements. We hypothesize that this robustness stems from modeling semantic descriptors within object masks and explicitly accounting for multi-view uncertainty, even in challenging scenes. Notably, in the 'Fr2_person' sequence with dynamic elements, the baseline degraded to below 0.5 MOTA and under 50% SR_{succ} at the 1 m threshold, exposing its limitation in filtering outliers caused by semantic label noise.



(a) Data Association Results (b) Pose Estimation Results

Fig. 4. **The qualitative results of data association and pose estimation.** (a) In each example, the left shows the query frame and the right shows the rendered landmark. Correct matches are drawn in green, while incorrect matches are shown in red. (b) The color bar illustrates the scale of translation error. Successful pose estimates appear as scatter points, color-coded by their error magnitude, and failed estimates are displayed as continuous gray lines.

Furthermore, as reported in Table II, both GOReloc and our method outperformed the feature point–based ORB-SLAM2 in terms of SR_{all}. This suggests improved scene invariance of semantic object–based approaches. In particular, our method shows substantial semantic discriminability, supporting robust and accurate pose estimation across the majority of frames.

3) Qualitative Evaluation: Fig. 4 qualitatively illustrates the effectiveness of the proposed approach in both data association and pose estimation. The results indicate that our method can effectively discriminate between visually similar objects and under complex object arrangements. This advantage can be attributed to our leveraging of the rich cues encoded in region-level semantic tokens and the incorporation of semantic variance tokens as informative priors.

TABLE III

ABLATION STUDY ON MODULES AND FOUNDATION MODELS

Dataset	Model	Module	Data Association		SR_{succ} [%]↑		
			F1↑	MOTA↑	@0.5m	@1m	@2m
Fr2_dishes	CLIP	Cos	**0.991**	**0.979**	**96.38**	**97.79**	98.79
		USS	0.981	0.952	93.20	95.95	**99.84**
	TAP	Cos	0.996	0.993	98.50	98.75	99.21
		USS	**0.998**	**0.996**	**98.99**	**99.07**	**99.40**
Fr2_desk	CLIP	Cos	**0.871**	**0.732**	**90.21**	**93.24**	**97.29**
		USS	0.861	0.714	88.11	91.37	96.24
	TAP	Cos	0.882	0.758	94.72	96.74	99.04
		USS	**0.887**	**0.768**	**96.47**	**97.20**	**99.17**
Fr2_person	CLIP	Cos	**0.882**	**0.742**	**67.62**	**72.59**	**77.40**
		USS	0.879	0.738	65.00	70.25	73.69
	TAP	Cos	**0.938**	**0.863**	81.24	86.29	87.98
		USS	0.933	0.854	**81.96**	**87.84**	**90.31**

C. Ablation Study

Table III reports the quantitative results across different foundation models and similarity modules. Within the M-STA framework, employing TAP tokens as semantic vectors allows the proposed USS to consistently achieve competitive performance across all sequences. This difference arises from feature extraction: unlike CLIP, which encodes both object and background, TAP extracts features solely from the masked object region. This leads to more distinct, background-free object descriptors. In particular, on 'Fr2_person', USS attains slightly lower F1 and MOTA but shows higher pose estimation performance. This is consistent with cross-session settings, where different object arrangements naturally reduce association metrics under dynamic factors. As illustrated in Fig. 4(a), our TAP-based method with the proposed USS demonstrates a significantly improved discriminative capability compared to the baseline and the cosine-similarity variant of our method.

V. CONCLUSION

In this work, we introduced *MUST-Loc*, a multi-view uncertainty-aware semantic token matching framework for global localization. Our method incrementally aggregates mean and variance tokens from multiple views and leverages the USS to establish reliable correspondences under semantic ambiguity, while estimating the camera pose by maximizing a Wasserstein alignment score between observations and projected landmarks. Experiments demonstrate that our method maintains robust localization performance under occlusions, noisy labels, and large-scale category diversity.

In future work, we will extend this framework to open-world localization by leveraging graph-based structures to handle cross-time variations, emphasizing robustness under temporal variations in object layouts.

REFERENCES

[1] M. Michalkiewicz, S. Bai, M. Baktashmotlagh, V. Jampani, and G. Balakrishnan, "Not all views are created equal: Analyzing viewpoint instabilities in vision foundation models," *arXiv preprint arXiv:2412.19920*, 2024.

[2] S. Matsuzaki, T. Sugino, K. Tanaka, Z. Sha, S. Nakaoka, S. Yoshizawa, and K. Shintani, "Clip-loc: Multi-modal landmark association for global localization

in object-based maps," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 13 673–13 679.
[3] S. Matsuzaki, K. Tanaka, and K. Shintani, "Clip-clique: Graph-based correspondence matching augmented by vision language models for object-based global localization," *IEEE Robotics and Automation Letters*, 2024.
[4] B. Zhou, Y. Meng, and F. Kai, "Object-based loop closure with directional histogram descriptor," in *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*. IEEE, 2022, pp. 1346–1351.
[5] Z. Qian, J. Fu, and J. Xiao, "Towards accurate loop closure detection in semantic slam with 3d semantic covisibility graphs," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2455–2462, 2022.
[6] Z. Cao, Q. Zhang, J. Guang, S. Wu, Z. Hu, and J. Liu, "Semantictopoloop: Semantic loop closure with 3d topological graph based on quadric-level object map," *IEEE Robotics and Automation Letters*, 2024.
[7] Y. Wu, Y. Zhang, D. Zhu, Z. Deng, W. Sun, X. Chen, and J. Zhang, "An object slam framework for association, mapping, and high-level tasks," *IEEE Transactions on Robotics*, vol. 39, no. 4, pp. 2912–2932, 2023.
[8] Y. Wang, C. Jiang, and X. Chen, "Goreloc: Graph-based object-level relocalization for visual slam," *IEEE Robotics and Automation Letters*, vol. 9, no. 10, pp. 8234–8241, 2024.
[9] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 573–580.
[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.
[11] X. Guo, J. Hu, J. Chen, F. Deng, and T. L. Lam, "Semantic histogram based graph matching for real-time multi-robot global localization in large scale environment," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 8349–8356, 2021.
[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy
[13] D. Shah, B. Osiński, S. Levine *et al.*, "Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action," in *Conference on robot learning*. PMLR, 2023, pp. 492–504.
[14] R. Mirjalili, M. Krawez, and W. Burgard, "Fm-loc: Using foundation models for improved vision-based localization," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 1381–1387.
[15] C. Kassab, M. Mattamala, L. Zhang, and M. Fallon, "Language-extended indoor slam (lexis): A versatile system for real-time visual scene understanding," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 15 988–15 994.
[16] T. Pan, L. Tang, X. Wang, and S. Shan, "Tokenize anything via prompting," in *European Conference on Computer Vision*. Springer, 2024, pp. 330–348.
[17] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.
[18] J. Wang, C. Xu, W. Yang, and L. Yu, "A normalized gaussian wasserstein distance for tiny object detection," *arXiv preprint arXiv:2110.13389*, 2021.
[19] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, pp. 1–10, 2008.
[20] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.