2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Context-Aware GAN-based Image Retrieval for Coarse Localization of Autonomous Robots

Ruphan Swaminathan and Pradyot Korupolu

***Abstract*— Effective localization is crucial for the reliable operation of autonomous delivery robots. This paper introduces ConLocGAN, a novel context-aware GAN, addressing challenges in Lidar-based localization. Our approach employs a two-step process, integrating image retrieval with Lidar-based localization. ConLocGAN extracts robust global descriptors for coarse pose estimator, which acts as a precursor for Lidar-based pose refinement. The discriminator in ConLocGAN identifies differences in images of the same scene under diverse conditions at the feature level. This information is then utilized to enhance localization-specific feature extraction by the generator in a self-supervised setting. Additionally, we present a simple data collection pipeline that is seamlessly integrated into routine robot operations. Using heatmaps for visualization, we demonstrate that our network learns robust descriptors by prioritizing static components of the scene while effectively disregarding environmental changes such as illumination and weather, as well as dynamic objects like people and vehicles. We further validate our method on the challenging CMU seasons dataset, where it outperforms state-of-the-art retrieval-based methods in coarse pose estimation.**

I. Introduction

Precise localization is pivotal for ensuring the smooth operation of autonomous delivery robots, preventing potential accidents such as falls from sidewalks and to navigate narrow pathways effectively. Our autonomous delivery robots are equipped with a comprehensive sensor suite, encompassing a 3D Lidar, Inertial Measurement Units (IMUs), and strategically positioned cameras, enabling precise localization within a pre-mapped environment. Fig. 1 showcases one version of our autonomous delivery robots capable of efficiently transporting food, beverages and packages in diverse indoor and outdoor settings. While Lidar-based localization stands out as a prevalent choice, addressing issues from other sensors like IMUs, encoders, and GPS errors, it is not without practical challenges. For instance, global pose initialization using Lidar incurs substantial computational overhead in large maps. Furthermore, the risk of delocalization emerges in environments lacking Lidar features, particularly in settings characterized by repetitive or uniform structures such as expansive open spaces, lengthy hallways, and corridors. Notably, these environments contain intricate semantic context, like signage, which plays a crucial role in mitigating delocalization. Recently, camera-based methods have gained traction, leveraging their unique advantages. While their accuracy may not match Lidar-based methods, they overcome drawbacks associated with other sensors and complement Lidar by providing rich color information, proving invaluable in the aforementioned environments.

The authors are with Ottonomy Inc, Santa Monica, California. {ruphan.s, pradyot.korupolu}@ottonomy.io

Fig. 1. Our autonomous last-mile delivery robot is designed to deliver food, beverages and packages in both indoor and outdoor settings. In the shown sample route, training images are collected and potential candidate matches are searched to train a context-aware image retrieval network.

Camera-based localization generally fall into three categories. Structure-based methods construct a 3D map of the environment during the mapping phase, associating 3D points with descriptors [1], [2]. In the localization phase, query image descriptors are matched with the 3D map, establishing 2D-3D correspondences used to estimate camera pose through algorithms like Perspective-n-Point combined with RANSAC. Image retrieval-based methods, on the other hand, create a descriptor database instead of a 3D map. During localization, an optimal search algorithm identifies the image from the database with the closest descriptor to the query image. Pose regressors utilize neural networks as a map, trained on images from the mapping phase, directly estimating the pose of a given query image [3], [4]. Each method possesses distinct advantages and disadvantages. Structure-based methods tend to be the most accurate among the three, albeit with higher memory and compute requirements. However, they may be less precise than Lidar-based methods due to probabilistic estimates for correspondences, particularly in environments with drastic changes. Pose regressors offer speed and a smaller memory footprint, yet they suffer from inaccuracies related to significant changes in viewpoint or lighting. Image retrieval, by itself, cannot suffice for localization even though it is robust to lighting and environmental changes as it only estimates camera pose based on the closest image in the mapping phase. To address these limitations, our approach combines image retrieval with Lidar-based localization, effectively leveraging the strengths

of both vision and Lidar-based methods. We utilize image retrieval as a coarse localization step. This serves as an initialization for a more fine-tuned Lidar-based localization process (out of scope for this work). The integrated approach aims to enhance the robustness and accuracy of last-mile delivery robots, enabling efficient and reliable operations in diverse real-world scenarios.

We adopt a two-step hierarchical process involving image retrieval for coarse localization, succeeded by Lidar-based pose refinement. In this paper we focus on the initial step, making the following key contributions:

- Introduction of a novel self-supervised GAN architecture for extracting robust global descriptors.
- Validation of the effectiveness of our approach across diverse settings, including indoor environments, a medium-scale university campus, and large-scale urban datasets.
- Demonstration of seamless integration into an autonomous delivery robot stack, spanning from data collection to deployment.

II. Related Works

A. Data Collection for Visual Localization

The acquisition of extensive datasets at minimal cost and time is pivotal for developing a robust visual localization system. Prior research has employed diverse strategies for efficient image-pose dataset collection. Barfoot et al. [5] leverage an autonomous path-following robot, utilizing a VT&R system to create a pose graph and sample pose labels. While the dataset spans over multiple seasons and time of the day, it rarely contains any dynamic objects in the scene such as pedestrians and vehicles. Other datasets, such as Google Street View Time Machine, utilize GPS information for capturing images from different viewpoints over time. A recent study [6] underscored the significance of accurate ground truth for training and evaluating camera-based localization algorithms. This work highlights challenges arising from different reference algorithms, like Structure-from-Motion (SfM), leading to varied local minima, influencing absolute pose accuracy evaluation.

Lidar-based localization using the LeGO-LOAM algorithm [7] offers accurate pose estimation, making it a potential ground truth source for vision-based localization algorithms. By measuring the uncertainty covariance of LeGO-LOAM, the successful localization of the robot can be assessed, and corresponding poses can be utilized as reliable ground truth for vision-based approaches. A similar approach has been proposed in [8], for an ICP based localization algorithm to demonstrate the potential of utilizing Lidar pose as groundtruth. To estimate the reliability of lidar-based pose for ground truth, we measure our robot's pose when its stationary in a dynamic environment with moving cars and people. In this environment, the obtained plots in Fig. 2 demonstrate the stability of lidar-based pose, with standard deviations of less than 2cm and 0.1 degrees for translational and rotational degrees of freedom, respectively. These results affirm the suitability of lidar-based pose as a dependable ground truth for vision-based localization algorithms.

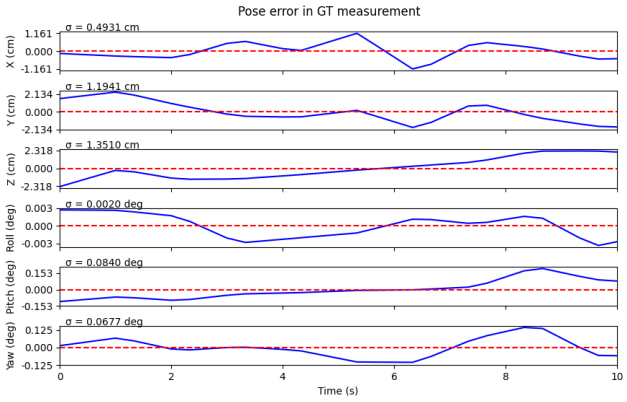

Fig. 2. Potential of Lidar pose as groundtruth for vision based coarse localization by measuring pose stability. When the robot is stationary, the standard deviations of translational and rotational degrees of freedom are less than 2cm and 0.1 degrees, respectively.

B. Image Retrieval for Visual Localization

NetVLAD [9] is a widely used technique that adopts a Convolutional Neural Network (CNN)-based end-to-end trainable model for descriptor extraction. It introduces a custom pooling layer that aggregates images from the same neighborhood into the same cluster. The novelty of NetVLAD lies in aggregating local descriptors from various regions of an image to generate a single global descriptor, resilient to changes in viewpoint, lighting, and partial occlusions. This resilience to changes in viewpoint, lighting, and partial occlusions is beneficial for structure-based localization methods, as noted in [1]. While multiple views improve 2D-3D correspondences, resilience to viewpoint changes may not always be advantageous, as indicated in a benchmark study [10]. In our specific camera-based localization application, the objective is to estimate a coarse pose, followed by fine-tuning to centimeter accuracy using an iterative lidar point cloud algorithms called FPFH [11]. For instance, relying solely on the closest retrieved image may lead to impractical outcomes when spatially distant images achieve higher matching scores, such as two images capturing a landmark from opposite sides. Consequently, it is imperative to tailor the retrieval algorithm to the application's needs rather than treating image retrieval as a standard approach for visual localization.

III. Proposed Method

In this section, we present ConLocGAN, a context-aware GAN designed to achieve coarse localization by retrieving the closest match based on the spatial context of the static components in a scene. Additionally, we outline our data collection pipeline that is seamlessly integrated into the daily operations of our autonomous robots.

A. Data Collection Pipeline

During the mapping phase, our autonomous robot is driven around the region of interest to gather diverse sensor data, including but not limited to Lidar scans and camera images.

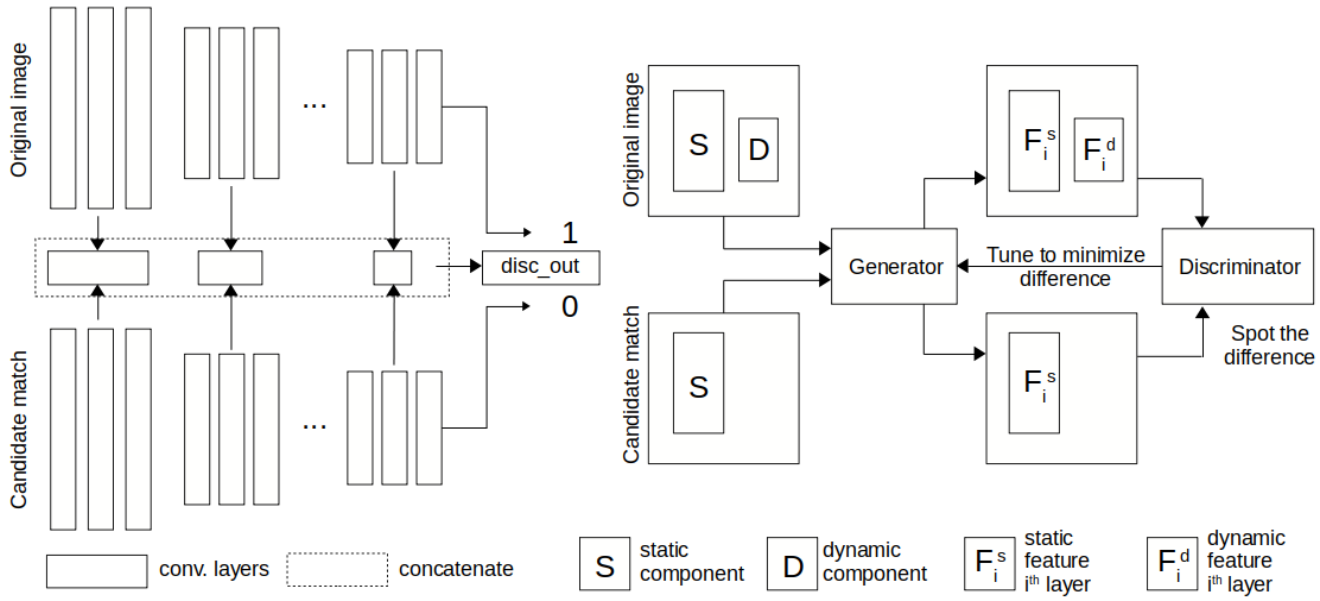

Fig. 3. Architecture of ConLocGAN. The generator is a Siamese network using an image and a random candidate match as input and generates a robust descriptor. The discriminator uses the intermediate feature maps to spot features of the dynamic components of the scene and aid the generator to learn localization-specific descriptor.

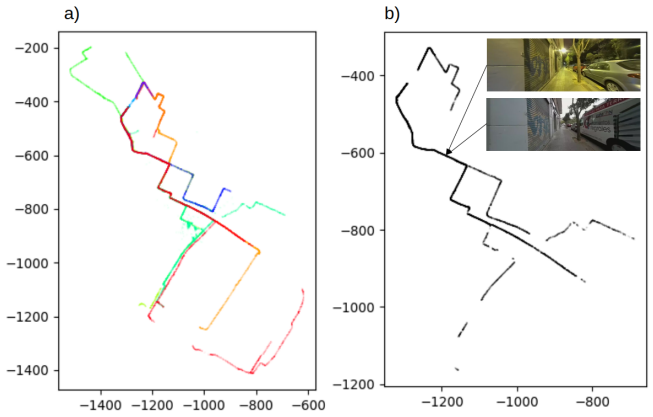

Fig. 4. Routes taken by the delivery robots. a) Each color represents the path taken by one robot during one day in a subset of the pre-mapped area. b) Images with at least one candidate match from a different color are chosen to minimize redundancy and learn context specific information.

Employing our proprietary SLAM algorithm, we construct a comprehensive 3D map of the operational area. Coarse visual localization, while not a primary requirement for autonomous navigation, serves to prevent manual interventions and scale up automation by automatically localizing the robot on boot-up and relocalizing during instances of delocalization. Thus, the image-pose dataset is collected during testing and delivery phase to minimize deployment delays. In the mapped environment, multiple robots make repeated trips between pickup and drop points in a day, as depicted in Fig. 4 a). Each color represents the traversals made by one robot in a single day. The image-pose dataset is curated by collecting and aggregating images with poses over a certain period.

For each collected image, we search for potential candidate images captured within a fixed distance threshold, denoted as d_λ. This value is determined based upon multiple factors such as width of the sidewalk/path taken by the robot, size of the map and the amount of overlap between traversals, typically falling in the range of 0.1 to 0.5 meters. During the training set creation, we use images with at least one candidate match, ignoring the remaining collected images, as illustrated in Fig. 4 b). Naturally, these images only cover a subset of the map and to create the retrieval database, we sparsely sample all the images collected during the mapping, testing and delivery phases to fully cover the region of interest. Aggregating images over different times of the day and various days, using different robots, ensures a diverse dataset without incurring significant cost or labor. This dataset encapsulates diverse conditions, including variations in illumination, environmental changes such as parked cars and changing billboards, and different camera parameters, including varying intrinsics. The simplicity and efficiency of our data collection system allows for repeated collections during different seasons if necessary, ensuring long-term localization accuracy.

B. ConLocGAN Architecture and Training

During each training iteration, an image I_{orig} from the training set is sampled along with a randomly chosen candidate match I_{cand}. The generator denoted by G takes in the image and its candidate match and returns their global descriptors along with intermediate feature maps as shown below:

$$Z_{org}, \{F^i_{orig}\}^{i_n}_{i=i_1} = G(I_{orig}) \quad (1)$$

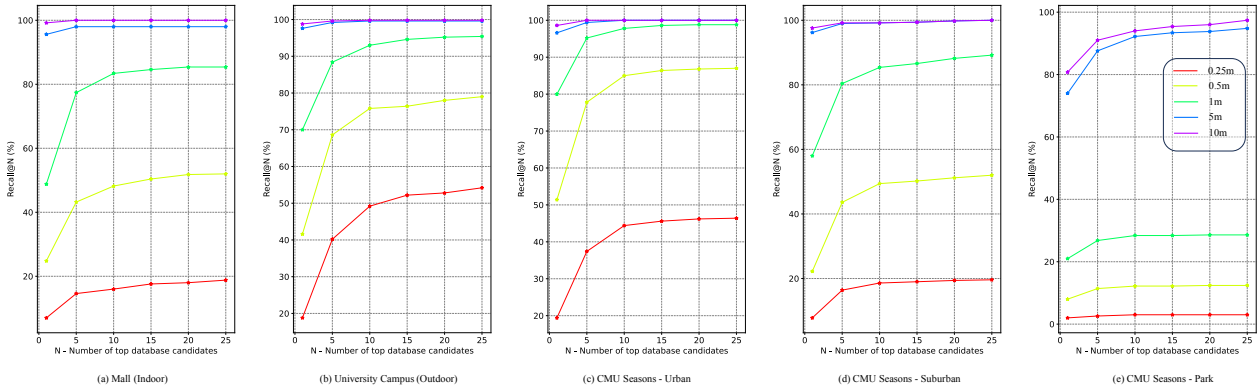

Fig. 5. Recall plots at different error thresholds for the mall, university campus and CMU seasons dataset. N refers to the number of closest candidates retrieved from the database and recall@N refers to the percentage of images with at least one correctly retrieved candidate within the error threshold.

$$Z_{cand}, \{F^i_{cand}\}^{i_n}_{i=i_1} = G(I_{cand}) \tag{2}$$

where $i \in \{i_1, i_2..i_n\}$ and i_n refers to the i^{th}_n layer of the generator. The n feature maps from the generator are passed to the discriminator D. The discriminator is composed of a set of $n+1$ residual blocks sequences. Each residual block sequence ψ_i is used to downsample the feature map F^i.

$$D(\{F^i\}^{i_n}_{i=i_1}) = \sigma\Big(\psi_{n+1}\big(\psi_1(F^1) \oplus \psi_2(F^2) \oplus \cdots \psi_n(F^n)\big)\Big) \tag{3}$$

The concept of employing a GAN model is drawn from the classic game of 'spot the difference.' At the image level, two significant differences exist between a training image and its candidate match. Firstly, dynamic components of the scene, such as people and vehicles, may vary between images. Secondly, variations in lighting conditions and seasonal changes can lead to differences in the illumination or texture of objects in the scene. Ideally, these differences should not persist at the descriptor level, as the global descriptor for an image and its candidate match should be the same. To enforce this, we employ a discriminator that identifies differences in the intermediate feature maps. This information is then fed back to the generator to enhance feature extraction, as illustrated in Fig. 3. Ideally, such a generator would produce localization-specific descriptors robust to environmental changes. The overall loss of the network is the weighted sum of three losses.

$$\mathcal{L}_{total} = \mathcal{L}_{self} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{desc}\mathcal{L}_{desc} \tag{4}$$

Eq. (5) to Eq. (9) show the formulation of the self-supervised loss, adversarial loss and the descriptor loss.

$$\mathcal{L}_{self} = -log\Big[\frac{f(Z_{orig}[0], Z_{cand}[0])}{\sum_j f(Z_{orig}[0], z_j)}\Big] \tag{5}$$

$$f(Z_1, Z_2) = \frac{Z_1.Z_2}{||Z_1||.||Z_2||} \tag{6}$$

$$z_j \in \big\{Z_{orig}[1 \cdots (N-1)] \oplus Z_{cand}[1 \cdots (N-1)]\big\} \tag{7}$$

$$\mathcal{L}_{adv} = \min_G \max_D \Big[log\big(D(\{F^i_{orig}\}^{i_n}_{i=i_1})\big) + log\big(1 - \big(D(\{F^i_{cand}\}^{i_n}_{i=i_1})\big)\Big] \tag{8}$$

$$\mathcal{L}_{desc} = \big|Z_{orig} - Z_{cand}\big| \tag{9}$$

where Z_{orig} and Z_{cand} are global descriptors with a batch size of N. Each residual block in the sequence ψ_i is followed by a max pooling layer. Thus, the number of residual blocks in the sequence is the number of times the input feature map is downsampled. Shallow feature maps have a bigger height and width but lesser channels than the deeper feature maps. To concatenate the outputs from the residual block sequences in Eq. (3), the height and widths must be identical. Thus, we have the following relation:

$$num\_blocks(\psi_1) > num\_blocks(\psi_2) > \cdots num\_blocks(\psi_n) \tag{10}$$

Apart from its use in downsampling, the higher number of residual blocks attending to the shallow feature maps were found to be effective to convergence of the network. This shows that learning better low level features contributes substantially to the overall effectiveness of the descriptor than learning to differentiate abstract features.

IV. Experimental Results

A. Implementation Details

The neural networks are implemented using PyTorch [12] and trained on an NVIDIA Tesla A100 GPU until convergence. We utilize an appropriate batch size to fully leverage the GPU memory and set the learning rate to 2e-5 with the Adam optimizer [13]. To help convergence, the learning rate is halved every 25 epochs. All networks process images at a resolution of 256x256 pixels except for the network on the indoor mall dataset which uses images at a resolution of 786x256 with a 360 degree view. We utilize the normalized

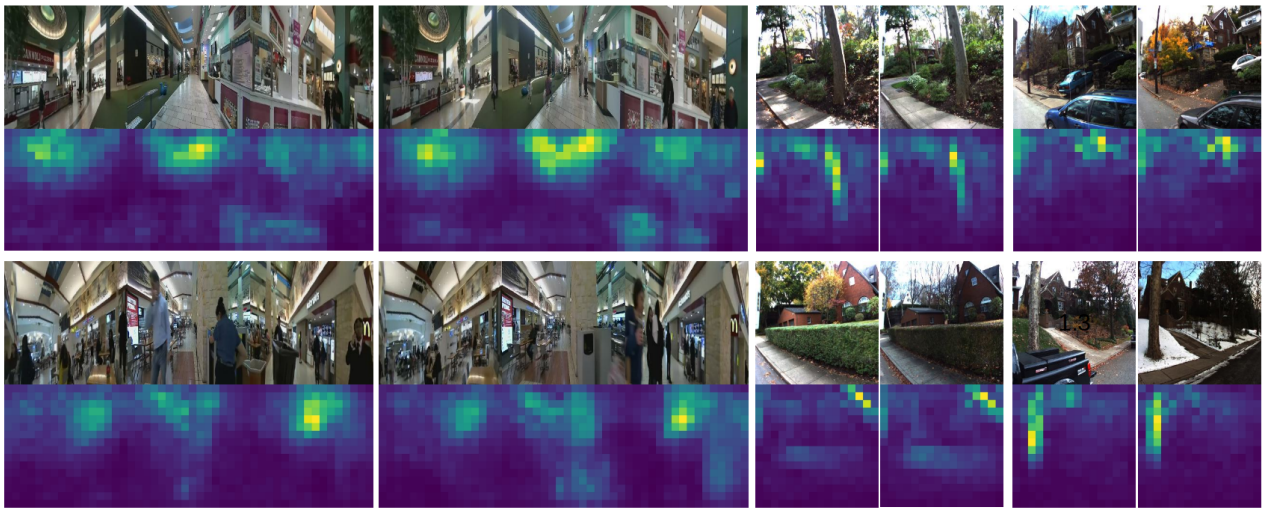

Fig. 6. Visualizing what has been learnt. Each image pair corresponds to a test image (unseen during training) and it's closest retrieved image. The heatmap visualization is based on the change in representation when occluded by a moving 16x16 patch. Combination of unique static components of the scene such as ceiling design, storefronts and beams in indoor environments and building silhouettes and trees in outdoor environments demonstrate context-aware retrieval. Dynamic components of the scene such as people and vehicles are ignored to ensure robust long-term localization.

TABLE I
RESULTS IN COMPARISON WITH EXISTING WORKS

Method	Urban(%) .25/.50/5.0 2/5/10	Suburban(%) .25/.50/5.0 2/5/10	Park(%) .25/.50/5.0 2/5/10
CSL [14]	**36.7**/42.0/53.1	8.6/11.7/21.1	**7.0**/9.6/17.0
FAB-MAP [15]	2.7/6.4/27.3	0.5/1.5/13.6	0.8/1.7/11.5
Xin et al. [16]	17.3/42.5/89.0	5.8/19.4/76.1	6.6/**23.1**/73.0
DIFL+FCL [17]	14.8/35.1/79.6	5.6/18.2/69.8	6.1/20.7/69.1
Wasabi [18]	6.6/23.1/73.0	3.8/13.9/67.3	2.4/9.1/54.5
DenseVLAD [19]	14.7/36.3/83.9	5.3/18.7/73.9	5.2/19.1/62.0
NetVLAD [9]	12.2/31.5/89.8	3.7/13.9/74.7	2.6/10.4/55.9
Ours	19.4/**51.4/96.6**	**10.2/27.6/96.4**	1.8/6.6/**73.8**

query descriptor to search through the database of descriptors generated during the training phase. The descriptor with the shortest L1 distance from the query is considered the closest match, and its associated pose serves as the coarse pose estimate for the query image.

B. Datasets

To comprehensively assess the robustness of our method, we conduct experiments using diverse datasets. Firstly, an indoor dataset is collected within a crowded mall environment, where our robot navigates through crowded spaces, picking up food from the food court, and delivering it to various locations within the mall with a delivery radius of about 250m. Additionally, an outdoor dataset is gathered within a university campus, featuring varying illumination conditions and seasonal changes, spanning daytime, dusktime, nighttime, and winter scenarios with a radius of about 500m. Furthermore, to benchmark and compare our method against existing literature, we train and evaluate on the extended CMU seasons dataset spanning much larger radius. This dataset is categorized into urban, suburban, and park settings, encompassing images captured over 12 different days throughout the year. The dataset incorporates environmental variations such as changing weather, foliage, illumination, and seasonal transitions. The inclusion of these challenging datasets ensures a thorough evaluation of our models across a diverse range of environmental conditions. For the indoor dataset, a distinct approach is taken, utilizing a larger stitched 360-degree view image. Unlike the outdoor datasets where the train and test data are captured by following a specific path, such as a sidewalk in the university campus dataset or on the roads in the CMU seasons dataset, there is no predefined path for indoor navigation within the mall. Opting for a 360-degree view, as opposed to relying solely on the front-facing camera, proves advantageous by having a holistic representation of the mall environment.

C. Results

We set the following hyperparameters in our experiments: λ_{adv} and λ_{desc} as 0.1 (in Eq. (4)), $n = 3$ and $i \in \{10, 22, 40\}$ (in Eq. (1) and Eq. (2)), and $num\_blocks(\psi_i)_{i=1}^{4} = \{4, 3, 2, 1\}$ (in Eq. (3)). Fig. 5 provides an overview of our model's performance through N versus recall@N plots, where N signifies the top N closest retrievals at various distance thresholds. Among these values, particular emphasis is placed on recall@1 at $d = 5m$. This is a critical evaluation metric for our application and refers to the percentage of images that were localized within 5 meters of the groundtruth location. While recall@N is better for higher N values, it necessitates the repetition of lidar-based pose refinement process for N times. However, this is not practical, making the recall@1 the primary focus from a deployment perspective. Further, achieving a near 100% coarse accuracy is preferred in comparison to performing better than other methods in lower distance thresholds as Lidar-based refinement improves pose accuracy. Additionally, dense sampling of the environment may not be feasible always, making

it challenging to guarantee a closest match within smaller distance thresholds like 0.25m or 0.5m for every possible query image.

Tab. I shows the comparison of our results with other works in the literature on the extended CMU seasons dataset. The results for the methods in comparison have been taken from an online benchmarking website [20]. Our method performs better than most of the works at a coarser scale achieving >95% accuracy in all datasets, except under the park setting of CMU seasons. This might be because of the high similarity in the images and lack of differentiating landmarks such as buildings. We observed a similar trend in our indoor mall dataset when using just the front camera images and was remedied by using the 360 view images to provide more context to the scene.

To gain insights into the inner workings of our network, we visualize the sensitivity to occlusions using a heatmap to understand the parts of the scene that contribute the most towards descriptor extraction. The heatmaps are generated by occluding a 16x16 square in the image and measuring the shift in the global descriptor. Fig. 6 showcases the heatmaps, highlighting the regions considered crucial for localization, such as patterns in the ceiling (indoors) and the silhouette of buildings (outdoors). It is evident that the network emphasizes on the static components of the scene and learns to ignore dynamic objects like people and vehicles.

V. Conclusions

In this study, we introduced ConLocGAN, a novel context-aware GAN designed for coarse pose estimation through image retrieval. Our motivation stems from addressing practical challenges encountered in past last-mile delivery projects, where limitations in lidar-based approaches posed issues in scalability of our autonomous robot operations. We showcased our data collection pipeline, seamlessly integrated into daily robot operations, facilitating the creation of diverse datasets—both indoors (crowded mall) and outdoors under varying illumination and weather conditions.

By evaluating our method against existing works on the challenging CMU seasons dataset, we showed that our method significantly outperforms other works in coarse pose estimation, a critical aspect for our specific application. The results and the heatmaps show the importance of large-scale and long-term visual localization in a self-supervised setting. The proposed method serves as a foundational step in our two-step hierarchical localization strategy, leveraging the strengths of diverse sensor modalities of cameras and lidar. We believe that our findings will inspire further research in this area, leading to the development of more practical and effective approaches for visual localization in challenging environments.

References

[1] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *CVPR*, 2019.

[2] P.-E. Sarlin, A. Unagar, M. Larsson, H. Germain, C. Toft, V. Larsson, M. Pollefeys, V. Lepetit, L. Hammarstrand, F. Kahl, and T. Sattler, "Back to the Feature: Learning Robust Camera Localization from Pixels to Pose," in *CVPR*, 2021.

[3] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015.

[4] S. Brahmbhatt, J. Gu, K. Kim, J. Hays, and J. Kautz, "Geometry-aware learning of maps for camera localization," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018.

[5] M. Gridseth and T. D. Barfoot, "Deepmel: Compiling visual multi-experience localization into a deep neural network," *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1674–1681, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:212628600

[6] E. Brachmann, M. Humenberger, C. Rother, and T. Sattler, "On the limits of pseudo ground truth in visual camera re-localisation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 6218–6228.

[7] T. Shan and B. Englot, "Lego-loam: Lightweight and ground-optimized lidar odometry and mapping on variable terrain," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 4758–4765.

[8] M. Ramezani, Y. Wang, M. Camurri, D. Wisth, M. Mattamala, and M. Fallon, "The newer college dataset: Handheld lidar, inertial and vision with ground truth," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020.

[9] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[10] N. Pion, M. Humenberger, G. Csurka, Y. Cabon, and T. Sattler, "Benchmarking image retrieval for visual localization," *CoRR*, vol. abs/2011.11946, 2020. [Online]. Available: https://arxiv.org/abs/2011.11946

[11] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (fpfh) for 3d registration," in *2009 IEEE International Conference on Robotics and Automation*, 2009, pp. 3212–3217.

[12] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[14] L. Svärm, O. Enqvist, F. Kahl, and M. Oskarsson, "City-scale localization for cameras with known vertical direction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 7, pp. 1455–1461, 2017.

[15] M. J. Cummins and P. Newman, "Fab-map: Probabilistic localization and mapping in the space of appearance," *The International Journal of Robotics Research*, vol. 27, pp. 647 – 665, 2008. [Online]. Available: https://api.semanticscholar.org/CorpusID:17969052

[16] Z. Xin, Y. Cai, T. Lu, X. Xing, S. Cai, J. Zhang, Y. Yang, and Y. Wang, "Localizing discriminative visual landmarks for place recognition," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 5979–5985.

[17] H. Hu, H. Wang, Z. Liu, C. Yang, W. Chen, and L. Xie, "Retrieval-based localization based on domain-invariant feature learning under changing environments," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 3684–3689.

[18] A. Benbihi, S. Arravechia, M. Geist, and C. Pradalier, "Image-based place recognition on bucolic environment across seasons from semantic edge description," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 3032–3038.

[19] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1808–1817.

[20] L. Hammarstrand. Long-term visual localization. [Online]. Available: https://www.visuallocalization.net