DeMixVPR: Integrating Depth and RGB Data for Robust Localization

Arda Eren Dogru
Politecnico di Torino
ardaeren.dogru@studenti.polito.it

Abstract

Visual geolocalization, or visual place recognition (VPR), aims to estimate the geographic location of an image by comparing it to a database of geo-tagged images. This work investigates various techniques to enhance the performance of VPR systems. Initially, extensive experiments are conducted on different miner and loss combinations using a sampled dataset. Promising combinations are then trained on the full dataset to identify the optimal configuration. Additionally, the impact of various optimizers and aggregation modules on the final performance is explored. Furthermore, depth information, obtained from depth maps generated by a pre-trained model, is integrated into the VPR pipeline. The proposed approach involves extracting features from RGB and depth images separately, concatenating the resultant feature maps, and subsequently processing them using a MixVPR module. A comparison is made between this approach and a state-of-the-art multimodal method, demonstrating the superior performance of the proposed technique. The source code is available at https://github.com/ardaerendogru/DeMixVPR

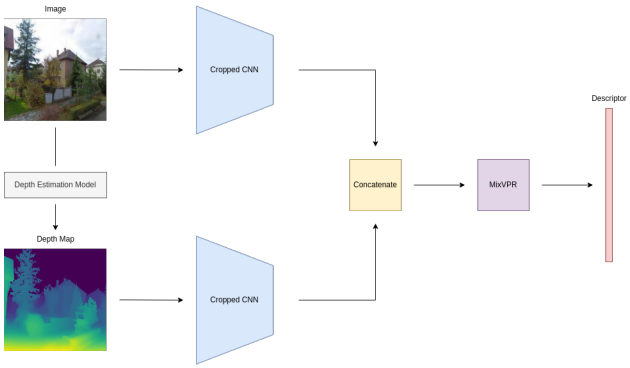

Figure 1. Illustration of the proposed multi-modal visual geo-localization (VG) framework, DeMixVPR, integrating monocular RGB images with depth information. The depth map is estimated from RGB input using a depth estimation model. Both RGB and depth data are processed through individual CNN encoders and subsequently concatenated. The concatenated features are passed through the MixVPR module to produce robust embeddings for geo-localization tasks.

1. Introduction

Visual geo-localization (VG), also referred to as Visual Place Recognition (VPR), is a fundamental task in computer vision, aiming to determine the geographic location of a given query image by matching it against a large database of geo-tagged images [3, 7, 14]. VG is crucial for various applications, including autonomous driving, augmented reality, and robotics, where accurate localization is essential for navigation and interaction with the environment [14, 18]. However, achieving robust geo-localization remains challenging due to diverse environmental conditions, variations in lighting and weather, and occlusions in urban settings.

Traditional VG methods primarily rely on RGB images, utilizing deep convolutional neural networks (CNNs) to extract discriminative features, which are subsequently used in image retrieval frameworks [19]. Despite demonstrating considerable success, these methods often encounter difficulties in conditions where visual cues are limited, such as low-light scenarios or environments with repetitive patterns. Recent research has begun exploring the use of multi-modal data, such as depth and LiDAR, to enhance the robustness of VG systems [5, 9–11, 15, 22].

LiDAR-based methods, including AdaFusion [10], have shown the effectiveness of integrating multiple data modalities for place recognition by dynamically fusing RGB and LiDAR features. While these approaches are promising, they typically require expensive LiDAR sensors and involve high computational costs due to the significant data dimensionality. Additionally, the lack of available depth data for many datasets, including those used in this project, limits the applicability of such methods [14]. To address these limitations, depth maps generated from monocular RGB images using the DepthAnything v2 model [21] are utilized in this work. Depth maps provide several advantages over LiDAR-based methods, offering a lightweight and cost-effective alternative that can be generated with any

RGB camera, making them accessible for a broader range of applications.

In this paper, a novel multi-modal VG framework is proposed that integrates both RGB and depth data to achieve robust localization under challenging conditions. The proposed model extracts features separately from RGB and depth data, concatenates their feature maps, and processes them using the MiXVPR module, a powerful aggregation technique that enhances feature robustness and discriminative power as shown in the figure 1. The effectiveness of the approach is demonstrated by comparing it against state-of-the-art models, such as DeAttVPR [11], which also utilizes generated depth data but employs an attention mechanism for feature extraction.

The main contributions of this paper are summarized as follows:

- A systematic grid search is conducted to optimize the choice of loss functions, miners, aggregators, optimizers, and learning rate schedulers, leading to a well-tuned model configuration that enhances performance across multiple datasets.
- The DepthAnything v2 model is employed to generate depth maps from monocular RGB images and extended current datasets, providing a cost-effective alternative to LiDAR-based approaches.
- A multi-modal VG model is introduced that effectively combines RGB and depth data, offering improved performance over single-modal models under diverse conditions.
- A comprehensive experimental evaluation is conducted on multiple datasets (GSV-XS, SF-XS, and Tokyo-XS), demonstrating that the proposed model outperforms existing state-of-the-art methods, including DeAttVPR.

By integrating depth data in a novel multi-modal framework and employing a thorough optimization process, it is aimed to advance the field of visual geo-localization, making it more robust and accessible for practical applications across various domains.

2. Related Work

Research in visual geo-localization (VPR) has primarily focused on leveraging image retrieval techniques, where a query image is matched against a large database of geo-tagged images to determine its location [3, 14, 19]. Traditional methods rely heavily on single-modal RGB data, employing deep convolutional neural networks (CNNs) to extract discriminative features that can differentiate between different places. NetVLAD [3], for instance, introduced a network architecture that integrates the VLAD layer into a CNN for robust place recognition. Similarly, various retrieval approaches have been explored, such as using view synthesis for 24/7 place recognition [19].

However, the performance of these methods can degrade significantly in challenging scenarios, such as low-light conditions or environments with repetitive patterns. To overcome these challenges, recent research has explored the use of multi-modal data. LiDAR-based approaches, such as AdaFusion [10], dynamically fuse RGB and LiDAR features to enhance place recognition robustness. However, these methods often require costly LiDAR sensors and entail high computational costs due to the high dimensionality of LiDAR data.

Moreover, several multi-modal approaches have recently emerged to further enhance visual place recognition. The work by Lyu et al. [13] introduces a multi-modal large language model (MLLM) approach that combines visual and language modalities to improve place recognition. This method, termed LLM-VPR, leverages vision-based retrieval to propose candidate locations and then uses language-based reasoning to refine the selection. The study demonstrates that combining the robust visual features produced by off-the-shelf vision foundation models with the reasoning capabilities of MLLMs can lead to more accurate place recognition without requiring VPR-specific supervised training.

In addition to these, the MSSPlace framework [15], which integrates visual and text semantics for multi-sensor place recognition, has shown that the use of text descriptions alongside visual cues can significantly improve localization performance. This method highlights the potential for combining multiple data types, including natural language, to address the complexities of diverse and dynamic environments.

Furthermore, recent studies have focused on enhancing feature extraction and aggregation methods for visual place recognition. Generalized Mean Pooling (GeM) [17] has become popular due to its lightweight design and strong performance in aggregating features for visual retrieval tasks.

MixVPR [2] introduces a holistic feature aggregation technique that uses feature maps from a pre-trained backbone to create a global descriptor. Unlike traditional methods like NetVLAD [3] or TransVPR [20], which rely on local or pyramidal aggregation, MixVPR utilizes a novel approach that combines global relationships between feature maps in a cascade of ”Feature-Mixer” blocks. These blocks consist of multi-layer perceptrons (MLPs) that integrate global features in a unified manner, eliminating the need for traditional local feature pooling strategies. This technique allows MixVPR to adapt to various visual contexts and scales, significantly improving visual place recognition performance. The method demonstrates its superiority by achieving state-of-the-art results across multiple

large-scale benchmarks while using fewer parameters and being faster than many existing methods.

In this work, the initial focus is on optimizing the network's performance through a systematic grid search, fine-tuning various components such as loss functions, mining strategies, optimizers, and learning rate schedulers to achieve a well-tuned baseline configuration. Building upon this optimized foundation, the study then introduces a novel multi-modal model that combines RGB and depth data. Unlike prior methods that rely on expensive sensors or are constrained by single-modal inputs, the proposed model utilizes depth maps generated from monocular RGB images using DepthAnything v2 [21]. This approach offers a cost-effective, scalable, and robust solution for visual geo-localization across diverse conditions, demonstrating superior performance compared to state-of-the-art methods, including DeAttVPR [11].

3. Methodology

The proposed visual geo-localization framework is designed to improve robustness and accuracy by integrating RGB and depth data through a multi-modal network. This section describes the experiment steps to achieve these objectives.

Baseline Model Selection. The initial phase involved experimenting with different pooling layers to establish a baseline RGB model. Average pooling and Generalized Mean (GeM) pooling [17] were evaluated by training the model on the GSV-XS dataset and validating on SF-XS validation dataset.

Loss Function and Miner Optimization. A grid search was performed on a sampled subset of the dataset to systematically explore various combinations of loss functions and mining strategies. The search encompassed several loss functions commonly employed in retrieval tasks, alongside multiple mining strategies designed to effectively select hard negative samples. The most promising combinations identified during this search were then used to train on the full dataset, optimizing the model's ability to learn discriminative features and improve performance.

Optimizer and Hyperparameter Tuning. Following the selection of the optimal loss and miner configuration, a second grid search was performed to identify the best optimizer, learning rate, and weight decay parameters. Various optimizers, including Adam, AdamW, ASGD and SGD, were considered. The grid search was conducted on sampled data, and the most effective combinations were applied to the full dataset.

Learning Rate Schedulers. After selecting the best optimizer configuration, different learning rate schedulers were tested to enhance model training stability and convergence. The Poly LR scheduler [6] and Cosine Annealing scheduler [12] were evaluated on the full dataset.

Advanced Feature Aggregation: MiXVPR. To further improve the model, the MixVPR [2] aggregation method was tested against the GeM pooling.

Development of Multi-Modal Network. The final stage of the methodology involved the development of a novel multi-modal model that leverages both RGB and depth data. Depth maps were generated using the DepthAnyting v2 model [21], a state-of-the-art depth estimation technique from monocular RGB images. Features were extracted separately from the RGB and depth inputs and concatenated before being fed into the MixVPR aggregation module. This approach allowed for the integration of complementary information from both modalities, enhancing the model's ability to handle diverse environmental conditions.

Comparison with State-of-the-Art Methods. The proposed multi-modal model was compared against DeAttVPR [11], which uses depth maps in an attention mechanism. The results demonstrated that the proposed model outperformed existing methods across all tested datasets, showcasing its effectiveness in both standard and challenging conditions.

4. Experiments and Results

This section presents the experimental setup and results obtained from the proposed visual geo-localization experiments. All experiments were conducted using PyTorch on a machine equipped with an NVIDIA RTX 4070 (Laptop) GPU, 64 GB of RAM, and an Intel i9-13900HX CPU.

4.1. Experimentation Details

Datasets. The training of the proposed visual geo-localization model was conducted on the GSV-XS dataset, a small version of the GSV-Cities dataset [1]. It consists of geo-tagged images from various urban environments, enabling the model to learn discriminative features across different locations.

For evaluation, two challenging benchmark datasets were employed: SF-XS and Tokyo-XS. The SF-XS dataset, a subset of the larger SF-XL dataset [4], was used for both validation and testing purposes. This dataset contains urban scenes from San Francisco with diverse viewpoint variations, complex architectural patterns, and dynamic lighting conditions, making it ideal for testing the robustness of geo-localization models under structured urban environments. The Tokyo-XS dataset, derived from the Tokyo 24/7 dataset [19], was employed exclusively for testing. Tokyo-XS features images captured at different times of day and in varying weather conditions, introducing additional challenges such as significant appearance changes, shadows, and occlusions. This dataset provides a rigorous testbed for evaluating model performance in dynamic, unstructured urban settings.

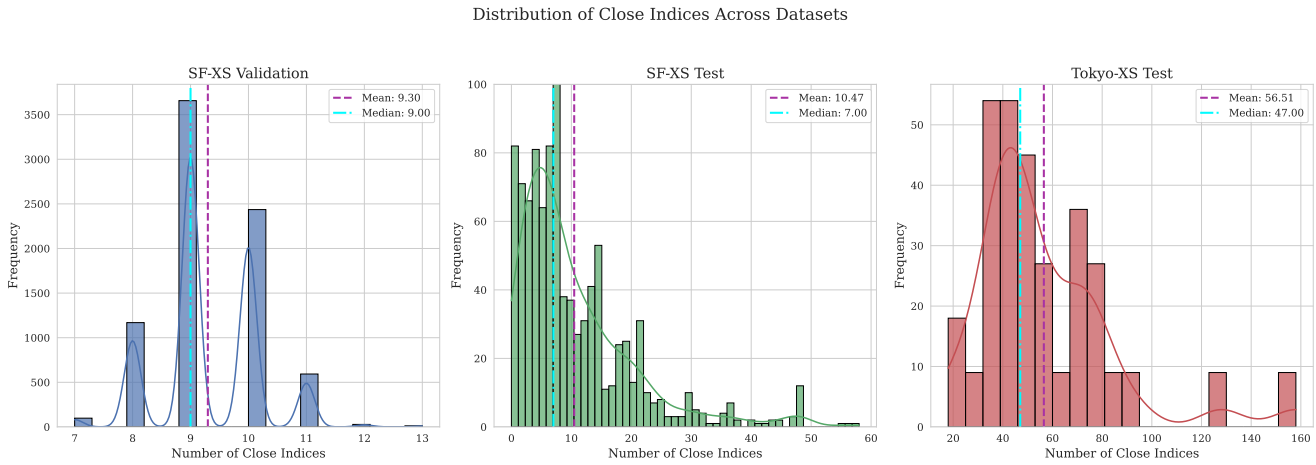

Figure 2. Distribution of Close Indices Across Datasets: SF-XS Validation, SF-XS Test, and Tokyo-XS Test. The figure illustrates the frequency distribution of the number of close indices (images are within a 25-meter radius) for each query across different datasets, highlighting the differences in query-reference match distributions that contribute to performance disparities.

Dataset Configuration. The experiments utilized both the full and sampled versions of the GSV-XS dataset. Except for the grid search optimization, which aimed to explore various loss functions, mining strategies, and hyperparameters, all experiments were conducted using the full dataset configuration. The full dataset comprised four images per place to ensure sufficient variability and context for learning discriminative features effectively. For the grid search, a subset of the GSV-XS dataset was used, containing approximately one-quarter of the distinct places and two images per place. This configuration enabled faster experimentation while maintaining a representative sample of the diversity present in the full dataset. The most promising combinations identified from the grid search were subsequently trained on the full dataset to confirm their effectiveness. All images are resized to 224x224.

Architecture. The proposed visual geo-localization model utilizes a multi-modal architecture that incorporates two separate ResNet-18 backbones: one for processing RGB images and another for processing depth maps. For initial experiments and optimization, a single ResNet-18 model was employed to simplify the process and focus on optimizing key parameters. The ResNet-18 network is truncated by removing all layers beyond the convolutional blocks to retain only the initial convolutional layers, which are essential for feature extraction relevant to the geo-localization task. To improve training efficiency and reduce computational costs, the first two convolutional blocks of the network are frozen, preventing their weights from being updated during training.

For the MixVPR configuration, the final convolutional block of ResNet-18 is also removed to produce feature maps of size $256 \times 14 \times 14$, as recommended in the MixVPR paper [2].

Training Configuration. Best models were saved using validation scores and used for testing. Different configurations designed to optimize the model's performance while balancing computational efficiency:

- **Baseline Selection:** The initial phase involved training models with *Average Pooling* and *GeM Pooling* to establish a baseline. Each model was trained for 30 epochs using an overfitting detector, which halted training early if no improvement was observed in the validation score for a predefined patience period. The training configuration employed Stochastic Gradient Descent (SGD) with a learning rate of 0.01, a weight decay of $1e-3$, and a momentum of 0.9. *Contrastive Loss* was used as the loss function for the baseline model training without a miner.

- **Grid Search Optimization:** For the grid search experiments, which aimed to explore different combinations of loss functions, mining strategies, optimizers, and learning rate schedulers, models were trained for 10 epochs. A pruning strategy was employed to terminate underperforming configurations early, thereby making the grid search process more efficient and conserving computational resources.

- **Refined Training Configuration:** After the grid search on optimizer, learning rate, and weight decay, it was found that the most promising configuration converged at the 27th epoch. Based on this observation, the number of training epochs was increased to 45 for

the subsequent experiments to ensure comprehensive learning.

Evaluation Protocol. The evaluation follows the standard recall@k metric commonly used in visual place recognition research [3, 14, 19]. A query image is considered successfully localized if at least one of the top-k retrieved reference images is within a 25-meter radius of the ground truth position. The recall@k scores were computed for k = 1 and 5, providing a comprehensive assessment of the model's retrieval performance. To efficiently perform similarity searches between query and database images, the FAISS (Facebook AI Similarity Search) library [8] was utilized. FAISS enables rapid computation of cosine similarities between global descriptors, ensuring scalability and high-speed retrieval even in large-scale datasets.

4.2. Baseline Selection

The initial phase of our experiments focused on establishing a strong baseline for the visual geo-localization model by evaluating different pooling strategies. Two pooling methods, *Average Pooling* and *Generalized Mean (GeM) Pooling* [17], were employed to determine their effectiveness in aggregating features. he results of these baseline experiments are summarized in Table 1.

Pooling Method	SF-XS (Val)		SF-XS (Test)		Tokyo-XS (Test)	
	R@1	R@5	R@1	R@5	R@1	R@5
Average Pooling	69.76	80.93	32.70	50.80	45.08	61.90
GeM Pooling	70.81	81.33	32.20	48.40	42.86	62.22

Table 1. Performance Comparison of Pooling Strategies

For the SF-XS test set, the lower recall scores can be attributed to the differences in query-reference match distributions between the validation and test sets. As shown in Figure 2, the SF-XS test set includes a considerable number of queries with few matches (1-5 matches: 28.2%, 0-1 matches: 8.2%), whereas the SF-XS validation set has no queries with fewer than 7 matches. This results in a more challenging retrieval task for the SF-XS test set, where 36.4% of the queries have minimal or no matching references. The variability in matches directly affects the model's ability to retrieve correct locations, explaining the drop in performance.

The Tokyo-XS test set presents additional challenges due to the significant appearance changes in images captured at different times of the day, including both day and night conditions. Given that the training dataset (GSV-XS) consists predominantly of day-time images, the model struggles to generalize to the drastically different conditions presented in the Tokyo-XS test set.

Overall, the results suggest that while the pooling strategies provide reasonable performance in controlled settings (SF-XS validation), their effectiveness diminishes under more varied and challenging conditions, such as those presented by the SF-XS and Tokyo-XS test sets. Despite these challenges, GeM pooling was selected as the baseline aggregation method due to its marginally better performance, likely resulting from its ability to capture both local and global patterns within the feature maps.

4.3. Loss Function and Miner Optimization

To optimize the model's performance for the visual geo-localization task, a systematic exploration of various combinations of loss functions and mining strategies was conducted. The goal was to identify the most effective combination that enhances the model's ability to learn discriminative features across different urban environments. The following loss functions were evaluated: *Angular Loss*, *Circle Loss*, *Contrastive Loss*, *MultiSimilarity Loss*, *SupCon Loss* (Supervised Contrastive Loss), *Triplet Margin Loss*, and *FastAP Loss*. Different mining strategies were also investigated, including *No Mining*, *Triplet Margin Miner*, *MultiSimilarity Miner*, *Pair Margin Miner*, *Distance Weighted Miner*, and *Batch Hard Miner* utilizing PyTorch Metric Learning library [16].

To ensure a diverse evaluation, most promising different combinations of loss functions and miners were selected and trained on the full dataset, avoiding repetition of similar configurations, and their performance was evaluated on the SF-XS validation dataset. The results are summarized in Table 2.

Miner	Loss Function	R@1	R@5
Batch Hard Miner	Contrastive Loss	72.84	83.00
Distance Weighted Miner	SupCon Loss	70.81	82.30
Pair Margin Miner	FastAP Loss	71.34	82.46
None	MultiSimilarity Loss	70.62	81.87

Table 2. Performance of Loss Function and Miner Combinations on SF-XS Validation Dataset

Based on the results, the combination of *Batch Hard Miner* with *Contrastive Loss* achieved the highest performance. Therefore, the combination of *Batch Hard Miner* with *Contrastive Loss* was selected for further experiments.

4.4. Optimizer, Learning Rate, and Weight Decay Optimization

To further enhance the model's performance, a grid search was conducted to explore various optimizers, learning rates, and weight decay values. The objective was to determine the optimal configuration that maximizes the model's recall scores on the SF-XS validation dataset. The optimizers evaluated included *SGD*, *Adam* , *AdamW* , and *ASGD*. The learning rates tested were: `1e-2`, `5e-3`, `1e-3`, `5e-4`, `1e-4`, `5e-5`, `1e-5`, and the weight decay values were: `0`, `1e-5`, `1e-4`, `1e-3`, `1e-2`.

Based on the preliminary evaluations, the most promising combinations of optimizers, learning rates, and weight decay values were selected and trained on the full dataset, avoiding repetition of similar configurations. The results of these configurations on the SF-XS validation dataset are summarized in Table 3. Notably, no configurations with *ASGD* were trained on the full dataset as they did not show promising results during initial evaluations.

Optimizer	Learning Rate (LR)	Weight Decay	R@1	R5
Adam	$1e-4$	$1e-5$	75.14	85.74
AdamW	$1e-4$	$1e-4$	75.03	85.25
SGD	$1e-2$	$1e-5$	73.98	84.22

Table 3. Performance of Different Optimizers, Learning Rates, and Weight Decay Combinations on SF-XS Validation Dataset

The combination of the *Adam* optimizer with a learning rate of `1e-4` and a weight decay of `1e-5` achieved the highest performance. This result indicates that the adaptive learning rate mechanism of Adam, which adjusts the learning rate for each parameter based on its historical gradient information, effectively balances convergence speed and model generalization.

The configuration using *AdamW* with a learning rate of `1e-4` and weight decay of `1e-4` showed slightly lower performance but still demonstrated strong results, suggesting that decoupling weight decay from the learning rate update can help prevent overfitting while maintaining stable convergence. However, the slight edge of Adam over AdamW in this context justifies its selection for further experimentation.

The *SGD* optimizer with a learning rate of `1e-2` and weight decay of `1e-5` also provided competitive results, but its performance lagged behind Adam and AdamW. Given that none of the configurations with *ASGD* showed promising results during initial evaluations, they were excluded from training on the full dataset.

Therefore, the configuration using the *Adam* optimizer with a learning rate of `1e-4` and a weight decay of `1e-5` was selected for further experimentation.

4.5. Learning Rate Scheduler Selection

After optimizing the choice of optimizer, learning rate, and weight decay, further experiments were conducted to determine the impact of different learning rate schedulers on the model's performance and convergence stability. Two widely used learning rate schedulers were evaluated: the *Poly LR scheduler* [6] and the *Cosine Annealing scheduler* [12].

Training was conducted for 45 epochs, based on prior observations that the *Adam* optimizer achieved convergence around the 27th epoch. The *Poly LR scheduler* achieved a Recall@1 of 77.46% and Recall@5 of 87.13% on the SF-XS validation dataset. In comparison, the *Cosine Annealing scheduler* resulted in slightly lower performance, with a Recall@1 of 74.99% and a Recall@5 of 85.39%. For a fair comparison, the best-performing model configuration from previous steps was also retrained for 45 epochs. This retraining resulted in a Recall@1 of 75.83% and a Recall@5 of 85.59%.

Given the higher recall scores achieved with the *Poly LR scheduler*, it was chosen for subsequent experiments.

Figure 3. This figure illustrates the query images alongside their prediction retrieved by the DeMixVPR model. The queries shown are those for which only the DeMixVPR model successfully retrieved the correct database images, while other models failed. This highlights the superior performance of DeMixVPR in these specific cases.

Method	Modality	Descriptor Size	SF-XS (Val)		SF-XS (Test)		Tokyo-XS (Test)	
			R@1	R@5	R@1	R@5	R@1	R@5
Optimized Baseline (GeM)	RGB	512	77.46	87.13	42.60	58.40	53.02	70.48
MixVPR	RGB	512	79.24	87.00	53.60	65.30	66.03	79.68
MixVPR	RGB	1024	80.23	87.84	55.30	66.80	67.62	80.95
Concatenation (GeM)	RGB + Depth	512	79.49	88.49	48.80	62.90	65.08	82.86
Concatenation (GeM)	RGB + Depth	1024	80.11	88.87	49.80	64.00	63.17	81.90
DeAttVPR	RGB + Depth	512	76.99	86.75	41.40	56.50	53.33	66.35
DeMixVPR (Ours)	RGB + Depth	512	82.42	89.72	57.20	**70.50**	**78.73**	88.25
DeMixVPR (Ours)	RGB + Depth	1024	**82.82**	**90.25**	**58.80**	68.30	76.19	**88.89**

Table 4. Performance comparison of different methods, descriptor sizes, and modalities on SF-XS validation, SF-XS test, and Tokyo-XS test datasets.

4.6. Novel Approach: DeMixVPR

Comparison of MixVPR and GeM Pooling Methods. The initial experiments focused on comparing the MixVPR aggregation method against the optimized GeM pooling strategy. As shown in Table 4, MixVPR was evaluated with descriptor sizes of 512 and 1024. The results indicate that MixVPR consistently outperforms GeM pooling across all datasets, particularly in challenging environments such as the Tokyo-XS test set. The improvement is evident when increasing the descriptor size, suggesting that MixVPR effectively captures both local and global contextual information, enhancing the model's ability to handle diverse visual scenes and scales. These findings align with the theoretical insights proposed in the MixVPR framework [2].

Development of the Depth-Enhanced Multi-Modal Network. Following the success of MixVPR, the study progressed to developing a multi-modal network that integrates both RGB and depth information. Depth maps were generated using the DepthAnything V2 model with pretrained base weights, aiming to enhance the feature representation by combining depth data with RGB input. The initial approach employed GeM pooling with concatenation of descriptors, inspired by the high-performing MSSPlace framework that utilizes a mix of LiDAR data, visual cues, and text semantics [15].

As shown in Table 4, while the concatenation method with GeM pooling delivered competitive results, the addition of depth information did not significantly outperform the RGB-only model using MixVPR. This indicates that although the concatenation approach with GeM pooling captures some complementary information, it may not fully leverage the potential advantages of incorporating depth data alongside RGB features.

Implementation and Evaluation of DeAttVPR. To further investigate the use of depth information, the DeAttVPR model was re-implemented and trained using the same optimized baseline settings. DeAttVPR incorporates an attention mechanism to dynamically fuse RGB and depth features, aiming to enhance the robustness of the geo-localization model. However, as shown in Table 4, the results were less favorable compared to the MixVPR approach. The performance gap observed between DeAttVPR and MixVPR suggests that while attention-based methods provide a sophisticated means of feature fusion, they may not be as effective in capturing the complex local and global feature relationships needed for challenging urban environments.

DeMixVPR. The proposed multi-modal approach, illustrated in Figure 1, leverages MixVPR for effective feature aggregation and demonstrates superior performance across all evaluated datasets. As shown in Table 4, DeMixVPR achieves the highest recall scores. This confirms the effectiveness of MixVPR in fusing RGB and depth data, resulting in a more robust representation that adapts well to diverse visual conditions.

Figure 3 further emphasizes the capability of the DeMixVPR model to handle complex challenges such as rural areas, structural changes, significant viewpoint variations, illumination changes, and occlusions. The figure presents examples of queries where only DeMixVPR successfully retrieved the correct database images, while other models failed. These results highlight the enhanced generalization capacity of DeMixVPR, which can learn richer and more discriminative feature representations than other methods.

While DeMixVPR shows substantial improvement over the baseline and other multi-modal approaches like DeAttVPR, certain challenges remain. The model's performance in environments with highly repetitive structures is still limited, indicating that future work could focus on further enhancing the model's ability to differentiate such visually similar yet geographically distinct locations. Overall, these findings underscore the strength of the DeMixVPR approach in integrating multi-modal data for robust and accurate visual geo-localization.

5. Conclusion

In this work, a novel approach to visual geo-localization was presented, focusing on enhancing performance through optimizing the model and leveraging multi-modal data. The

effectiveness of the proposed DeMixVPR model was comprehensively evaluated across various datasets, including the challenging Tokyo-XS test set, which encompasses diverse day and night scenes. The empirical results clearly demonstrate that DeMixVPR, which integrates depth information with RGB data and employs MixVPR for effective feature aggregation, consistently outperforms state-of-the-art methods like DeAttVPR.

The grid search optimization process contributed significantly to identifying the best configuration of loss function, miner, optimizer, and learning rate scheduler, resulting in a robust baseline model. This optimized model provided a strong foundation for integrating depth information and developing the multi-modal DeMixVPR model.

While DeMixVPR exhibits promising performance, certain areas warrant further investigation. In future work, advanced fusion techniques, such as attention mechanisms or cross-modal learning strategies, could be explored to capture more intricate relationships between modalities. Moreover, while the model performs well across various challenges, including rural areas, structural changes, significant viewpoint variations, illumination changes, and occlusions, addressing limitations in scenarios with highly repetitive structures remains a crucial avenue for further exploration.

In conclusion, this research underscores the potential of multi-modal approaches in achieving more accurate and robust visual geo-localization. By effectively combining RGB and depth data and leveraging advanced feature aggregation techniques, it is believed that the proposed DeMixVPR model presents a significant advancement in visual place recognition, particularly in challenging urban environments. The outcomes of this study are expected to contribute to the ongoing efforts to improve visual geo-localization and make it more accessible for various applications, ranging from autonomous navigation to augmented reality.

References

[1] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. Gsv-cities: Toward appropriate supervised visual place recognition. 2022. 3

[2] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. Mixvpr: Feature mixing for visual place recognition, 2023. 2, 3, 4, 7

[3] Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition, 2015. 1, 2, 5

[4] Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications, 2022. 3

[5] Fabien Bonardi, Samia Ainouz, Rémi Boutteau, Yohan Dupuis, Xavier Savatier, and Pascal Vasseur. Phrog: A multimodal feature for place recognition. *Sensors*, 17(5), 2017. 1

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation, 2017. 3, 6

[7] Stephen Hausler, Adam Jacobson, and Michael Milford. Multi-process fusion: Visual place recognition using multiple image processing methods. 2019. 1

[8] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus, 2017. 5

[9] Giseop Kim, Yeong Sang Park, Younghun Cho, Jinyong Jeong, and Ayoung Kim. Mulran: Multimodal range dataset for urban place recognition. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6246–6253, 2020. 1

[10] Haowen Lai, Peng Yin, and Sebastian Scherer. Adafusion: Visual-lidar fusion with adaptive weights for place recognition, 2021. 1, 2

[11] Fang Li, Mingyu Zhou, and Wei Liu. Deattvpr: Depth-attention visual place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 3

[12] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2016. 3, 6

[13] Zonglin Lyu, Juexiao Zhang, Mingxuan Lu, Yiming Li, and Chen Feng. Tell me where you are: Multimodal llms meet place recognition, 2024. 2

[14] Carlo Masone and Barbara Caputo. A survey on deep visual place recognition. *IEEE Access*, 9:19516–19547, 2021. 1, 2, 5

[15] Alexander Melekhin, Dmitry Yudin, Ilia Petryashin, and Vitaly Bezuglyj. Mssplace: Multi-sensor place recognition with visual and text semantics, 2024. 1, 2, 7

[16] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. Pytorch metric learning, 2020. 5

[17] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1655–1668, 2019. 2, 3, 5

[18] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale, 2018. 1

[19] Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1, 2, 3, 5

[20] Ruotong Wang, Yanqing Shen, Weiliang Zuo, Sanping Zhou, and Nanning Zheng. Transvpr: Transformer-based place recognition with multi-level attention aggregation, 2022. 2

[21] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2, 2024. 1, 3

[22] Fangwen Yu, Yujie Wu, Songchen Ma, Mingkun Xu, Hongyi Li, Huanyu Qu, Chenhang Song, Taoyi Wang, Rong Zhao, and Luping Shi. Brain-inspired multimodal hybrid neural network for robot place recognition. *Science Robotics*, 8(78):eabm6996, 2023. 1