# Exploring the Addition of Attention Mechanisms guided by Semantic Segmentation in Place Recognition

Gregory Holder
*Networking and Telecommunications*
*Université de Technologie de Troyes*
Troyes, France
g.c.holder@student.utwente.nl

Prawin Kumar Srinivasa Kumar
*Masters Embedded Systems*
*University of Twente*
Enschede,The Netherlands
p.k.srinivasakumar@student.utwente.nl

Christoforos Tsiolakis
*School of Electrical and Computer Engineering*
*Aristotle University of Thessaloniki*
Thessaloniki, Greece
c.tsiolakis@student.utwente.nl
tsiolakis@ece.auth.gr

Christos Siantsis
*School of Electrical and Computer Engineering*
*Aristotle University of Thessaloniki*
Thessaloniki, Greece
c.siantsis@student.utwente.nl
chrisian@ece.auth.gr

***Abstract*—In this paper, we investigate the integration of attention mechanisms guided by semantic segmentation into a VPR network which uses the Generalized Constructive Loss (GCL) function. Our objective was to identify the impact of this approach on the network's performance and to assess its effectiveness in enhancing the overall network. To derive informative features from input images, we employ the DeepLabv3+ model. We take two distinct approaches: the first involves generating a mask that highlights significant object categories crucial for VPR, such as stationary objects, while the second approach incorporates all semantic features as input to the VPR network, with their weighting determined by the network itself. While our results may have fallen slightly below the benchmark, our findings shed light on the potential avenues for improvement and offer valuable insights into the utilization of attention mechanisms in VPR systems.**

## I. Introduction

Visual Place Recognition (VPR) plays a crucial role in various fields, such as robotics, where long-term autonomy and visual sensing are required. Recent years have witnessed an increasing focus on VPR research due to its significance. A common approach to assess image similarity is to employ Convolutional Neural Networks (CNNs) for extracting relevant image features. These features are represented as vectors and their distances are used to determine the similarity between images. However, a challenge arises from the presence of dynamic objects in the images, including cars, people, and the sky, which may introduce noise and hinder accurate place recognition. This project aims to address this challenge by mitigating the impact of dynamic objects on the Generalized Constructive Loss (GCL) network. In [1], a graded similarity approach is proposed, which assigns continuous labels to image pairs based on metadata, reducing noise and training complexity. We propose to combine this training approach with semantic segmentation to transform the input. Specifically, we explore the implementation of an attention mask that focuses on static objects in the image, in addition to augmenting the input labels of the dataset through separate experiments.

## II. Related Work

### A. Semantic Segmentation

Paoloceli et al. [2] designed a dynamic attention-based mechanism that focuses on place recognition tasks by virtue of semantic segmentation. It harnesses multi-scale information about visual appearance and semantic content to generate global descriptors by leveraging features located at multiple scales that can efficiently capture discriminative objects of different sizes located in an urban setting. Contrary to multiple works performed previously where semantic information was used in a top-down approach, this work involved using a data-driven approach wherein the model determined relevant semantic information by allowing the VPR to guide the semantic segmentation network. For training, a synthetic dataset was created and adversarial training was used to reduce the gap between the synthetic and the real-world images. Furthermore, the authors identified that the model generalized better to unknown data. This paper displays similarities to our approach since it utilizes semantic segmentation and combines it with a VPR network. In our implementation though, the semantic information deemed more useful is predetermined and used only in the network's input.

In [3], DeepLabV3+ is presented as an improved semantic segmentation network over DeepLabV3 [4]. Atrous convolution is used in order to expand the receptive field of the filters

and incorporate multi-scale context. DeepLabV3+ utilizes this architecture and builds on top of it for the encoder module. The low-level features from the ResNet-101 which was employed as the network backbone are concatenated with features from the encoder having the same spatial resolution. The performance of the decoder is improved by reducing the number of channels of the low-level features and feature refinement. This network architecture was trained in the Cityscapes dataset [5] showing a significant improvement over the baseline.

Peng et al. [6] created a novel architecture SRALNet which proved that adding attention mechanisms into feature embedding can provide superior image representation performance. This involved adding spatial activations from the feature maps of the last convolutional layer are treated as deep local features. The local features are then divided into clusters. Irrelevant features in each cluster are suppressed using a locally weighted approach to its feature distribution which in turn reallocates the necessary ones into sub-clusters. They also employ semantic-constrained initialization in order to facilitate partitioning the encoding space using semantic priors. By these techniques, the authors identified that local attention uses a mutually beneficial training mechanism between data-driven learning and semantic priors.

Choi et al. [7] developed a novel de-attention mechanism that allows dynamic clutters to be excluded from their local features. Existing data-driven attention mechanism models make use of pre-trained ImageNet weights that focuses on non-essential objects which can hinder the performance of Image-retrieval tasks. This method employs a technique of configurable semantic guidance performed by the user to develop a de-attention layer trained in combination with Sharpened Triplet Marginal Loss (STMP) loss that would besides aiding better scene recognition also enhance contrastive learning through better discrimination between positive and negative samples. Deriving inspiration from this work we identified that it is much more efficient to inform the network things not to pay attention than vice-versa. We feel this work is the closest to our implementation.

### B. Attention mechanism

Park et al. [8] proposed a novel architecture named Bottleneck Attention Module (BAM) which was an efficient way of improving the network's representational power which produces an attention map emphasizing regions of interest. By exploiting the interchannel relationship from each channel and a channel vector that subtly embeds global information within each channel is obtained by pooling the feature maps. The spatial information is obtained by performing dilated convolutions to enlarge the receptive field in order to prioritize contextual information and the proposed bottleneck backbone is added to this branch as it minimizes the number of parameters and computational cost. This network is accordingly scaled so this can be combined with channel attention to creating a 3D attention map through element-wise summation for the most optimal performance.

CBAM [9] works more similarly to BAM and uses the BAM module at every convolutional block instead of the bottleneck of the network like the latter. From an intermediate feature map, CMAP infers a 2D spatial attention map that leverages the inter-spatial relationship of features to identify areas on the images that it is supposed to focus on. Additionally, a 1D channel attention map incorporates using both average and max-pooled features to obtain fine-grained channel-wise attention. A complementary network that focuses on "what" and "where" are crucial for paying attention to, which are then sequentially arranged to obtain an overall module that induces the model to focus on the right objects.

Wang et al. [10] proposed a residual attention network that employs multiple attention modules that can be scaled to synthesize an end-to-end trainable deep network in a single feedforward process by utilizing the top-down attention feedback and a corresponding bottom-up feedforward backbone that embeds soft weights on features to make them more attention-aware. Wherein each trunk learns attention specific to its features. The stacked attention modules supplement the residual learning and successively fine-tune the feature maps. Based on the global information obtained from the feedforward network this is then combined with the feature maps from the original image to obtain better performance.

## III. Model Implementation

### A. Semantic Segmentation

In this work, semantic segmentation is employed to extract detailed object information from input images. The DeepLabv3+ model [3], a state-of-the-art architecture for semantic segmentation, is utilized for this purpose. The model used features from a ResNet101 backbone. DeepLabV3+ is a well-established architecture widely used for semantic segmentation tasks. It was pretrained on the Cityscapes dataset [5], which consists of high-resolution urban street scenes, making it particularly suitable for capturing intricate object boundaries and fine-grained semantic information which are useful for VPR tasks. On applying the DeepLabv3+ model to the input images, we obtain 19 semantic activation channels corresponding to the following objects:

| | | | |
|---|---|---|---|
| **road** | **sidewalk** | **building** | **wall** |
| **fence** | **pole** | **traffic light** | **traffic sign** |
| **vegetation** | **terrain** | **sky** | **person** |
| **rider** | **car** | **truck** | **bus** |
| **train** | **motorcycle** | **bicycle** | |

Each channel represents a matrix of size $H \times W$, where $H$ and $W$ denote the height and width of the input image, respectively. These matrices contain information indicating the likelihood of each pixel belonging to the corresponding object category. Initially, we attempted to incorporate the DeepLabV3+ network directly into our model. However, this proved to be quite challenging as this resulted in significantly increased training times. To address this issue, we pre-generated the semantic activation matrices. The values of the matrices were normalized to the range [0, 255] and cached as grayscale images of size $H \times W$, allowing for quick retrieval

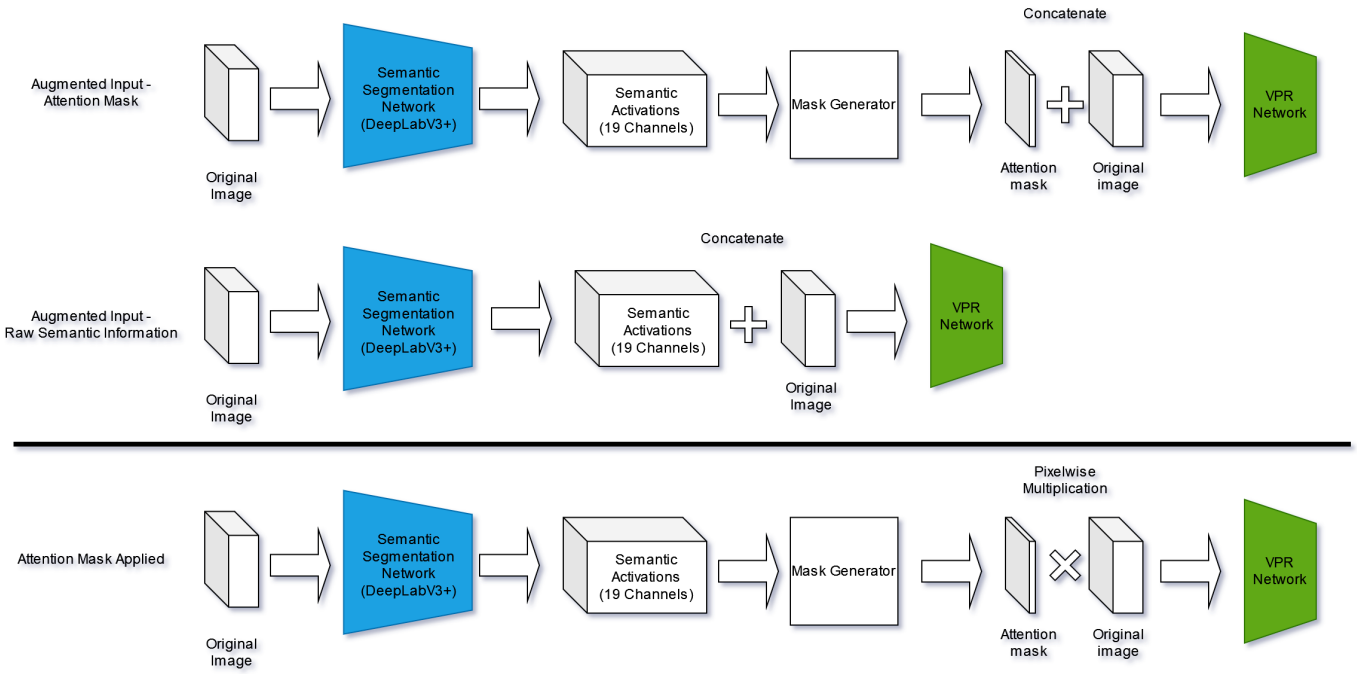


Fig. 1: Block Diagram of the Implemented Architecture. Top: Augmented Input - Attention Mask. The attention mask, generated by averaging the semantic activation channels related to non-time-varying objects, is directly concatenated with the input image. Middle: Augmented Input - Raw Semantic Information. The 19 semantic activation channels, corresponding to different object categories, are directly concatenated with the input image. Bottom: Attention Mask Applied with Pixelwise Multiplication. The attention mask obtained from the semantic activation channels is applied to the input image using pixelwise multiplication, highlighting the regions crucial for visual place recognition.

and utilization, allowing them to be efficiently used as inputs for our experiments.

### B. Attention Mask

In our model implementation, we leveraged the semantic activation channels obtained from the DeepLabV3+ network to generate an attention mask aimed at enhancing the performance of the VPR Network. The objective of the attention mask was to emphasize the crucial regions of an image for VPR tasks by specifically focusing on static objects such as buildings, traffic signs, sidewalks, and other non-time-varying elements instead of non-static objects such as pedestrians or vehicles. After trial and error we determined that the following object labels were most relevant for generating the attention mask:

| | | | |
|---|---|---|---|
| **road** | **sidewalk** | **building** | **wall** |
| **fence** | **pole** | **traffic light** | **traffic sign** |
| **vegetation** | **terrain** | | |

The above object labels all correspond to static objects in the image (time-invariant). To create the attention mask, we employed a simple yet effective approach of averaging the values of these selected semantic activation channels. This aggregation process resulted in a comprehensive attention mask that highlighted the significant regions indicative of specific locations.

### C. Integration with the VPR Network

In our efforts to improve the performance of the Visual Place Recognition (VPR) network, we investigated three different approaches that utilized the semantic activations and attention mask obtained from the DeepLabV3+ model. The first two approaches focused on augmenting the input by incorporating the semantic activations or attention mask directly into the input pipeline of the VPR network. However, both approaches presented certain challenges and limitations.

The initial approach involved concatenating the attention mask with the RGB image, however in order to accommodate the four channels (RGB+Mask), we had to alter the input layer of the VPR network. While this approach seemed promising, it suffered from information loss due to the averaging process used during the generation of the attention mask. Moreover, the modification of the input layer resulted in the loss of the pre-trained weights, as initially employed in [1].

The second approach attempted to address the information loss issue by concatenating all 19 semantic activation channels with the RGB image, thereby expanding the input layer to 22 channels (RGB+19 semantic activations). While this approach preserved the complete semantic information, it encountered challenges stemming from the large dimensionality of the input. The increased complexity of the input data seemed to confuse the network, hindering its ability to effectively learn

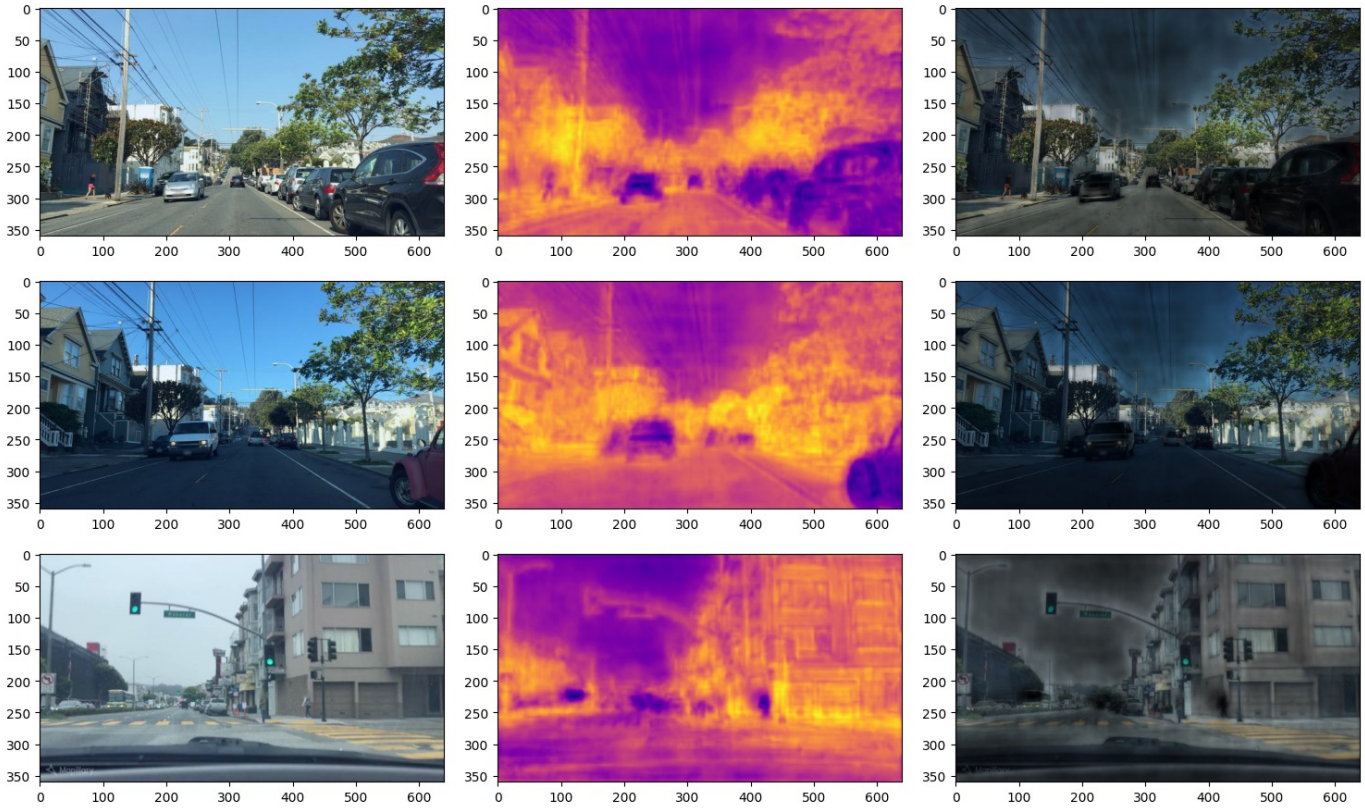

Fig. 2: Left: Three Images from the MSLS Reduced Dataset. Middle: Their Attention Masks. Right: The Images having applied the Masks

and extract meaningful location-specific features. Similar to the first approach, the redefinition of the input layer led to the loss of the pre-trained weights, further diminishing the overall performance of the VPR network.

To overcome the limitations of the previous approaches, we devised a more effective strategy, the pixel-wise multiplication of the attention mask with the RGB image. This approach circumvented the need to modify the input layer, thereby retaining the pre-trained weights and leveraging the full capability of the VPR network. By applying the attention mask in a pixel-wise manner, we were able to selectively emphasize the regions deemed crucial for location recognition, effectively filtering out irrelevant visual elements. This technique proved to be the most successful with reference to the VPR network's performance against the other approaches, as it preserved both the original information of the RGB image and the discriminative power of the attention mask. To provide visual insight into the application of the attention mask, we present examples in Figure 2, illustrating the pixel-wise multiplication of the attention mask with representative images. These examples serve to demonstrate the ability of the attention mask to highlight and emphasize the significant spatial regions essential for accurate place recognition (examples in Figure 2 include the highlighted traffic lights, signs etc), thereby enabling the VPR network to focus on salient and stationary features while disregarding irrelevant and dynamic elements.

## IV. Experiments and Results

To evaluate the effectiveness of the proposed approaches, we conducted a series of experiments using a subset of the Mapillary Street-Level Sequence dataset [11] (which was also used in [1]) where the following cities were included for training:

**amman boston london manila zurich**

While for validation:

**cph sf**

The smaller subset was used instead of the entire MSLS dataset to speed up training and validation times and decrease the required computational needs.

### A. ResNet50 backbone with modified input shape

In line with our design, we attempted to change the input shape of the network, replacing the original $3 \times H \times W$ input tensor with either a $4 \times H \times W$ tensor where the original image and the mask are concatenated, or a $22 \times H \times W$ tensor in which the image and all 19 semantic activations are concatenated. The fundamental idea behind both approaches would be to either augment the model with information about the regions of interest within the image(using the generated attention mask) or augment the model with per-pixel semantic context. As both approaches involve changing the input shape, this requires replacing the first convolution layer with one with the matching input channels. An unfortunate consequence

TABLE I: Performance Metrics

| Method | all_recall@1 | all_recall@5 | all_recall@10 | all_recall@20 | all_map@1 | all_map@5 | all_map@10 | all_map@20 |
|---|---|---|---|---|---|---|---|---|
| Normal-0-2 | 0.604 | 0.747 | 0.789 | 0.838 | 0.604 | 0.402 | 0.347 | 0.333 |
| Normal-2-2 | 0.619 | 0.741 | 0.796 | 0.836 | 0.619 | 0.408 | 0.351 | 0.339 |
| Normal-0-4 | 0.611 | 0.755 | 0.800 | 0.853 | 0.611 | 0.413 | 0.357 | 0.344 |
| Normal All | 0.611 | 0.762 | 0.785 | 0.842 | 0.635 | 0.420 | 0.360 | 0.345 |
| Masked-0-2 | 0.526 | 0.682 | 0.739 | 0.776 | 0.526 | 0.337 | 0.286 | 0.273 |
| Masked-2-2 | 0.542 | 0.705 | 0.750 | 0.804 | 0.542 | 0.352 | 0.300 | 0.287 |
| Masked-0-4 | 0.554 | 0.699 | 0.749 | 0.809 | 0.554 | 0.351 | 0.300 | 0.286 |
| Masked All | 0.561 | 0.712 | 0.754 | 0.795 | 0.561 | 0.359 | 0.309 | 0.296 |

Note: X and Y in "Normal-X-Y" and "Masked-X-Y" refer to the number of input and final layers of the ResNet50 backbone which are trained during each experiment correspondingly.

of this is the pre-trained weights for the first layer of the ResNet50 model getting lost. As a result, our attempts at modifying the input shape by replacing the first layer so far have proven ineffective rendering the overall performance of the network below the benchmark.

### *B. Unmodified ResNet50 backbone*

Firstly, the baseline model was finetuned where only the last two layers were included for training the model keeping the remaining layers frozen as proposed in [1]. Additionally, we wanted to test the impact of applying our pre-generated masks to the input images, again only training the last two layers of the backbone. With this approach, we do not change the input shape of the network, instead, we attempt to encode the attention mask into the input images directly using pixel-wise multiplication.

Considering that we are now giving the backbone images which have been altered, we theorized that it may be beneficial to also train the first two convolutional layers aiming to better integrate the weighted images. For the sake of comparison, we also tried this same first-two and last-two training approach with the original unmodified images. Furthermore, we tried training the last four layers, with both the original images and images with the mask applied. This was performed to test if training more layers in addition to preserving the core of the backbone would yield better results instead of interfering with the initial stages of the network. Finally, after observing that increasing the number of fine-tuned layers reduced the performance gap between the normal and masked runs, we decided to test the impact of fine-tuning the entire ResNet 50 backbone.

As per the results in Table I, the runs in which the attention mask was applied always performed worse than the baselines without the masks. However, while the baseline runs only slightly improved when more layers were finetuned, the performance gap between runs with and without masking decreased when more layers were left unfrozen. This would suggest that finetuning more of the ResNet50 backbone can yield better results.

## V. Discussions and Conclusion

In conclusion, our study investigated the incorporation of semantic segmentation information into a Visual Place Recognition (VPR) network to enhance its performance. Through various experiments, we explored different approaches to utilize semantic activation channels and an attention mask derived from the DeepLabV3+ model. Our findings demonstrated that the application of an attention mask, obtained through pixel-wise multiplication with the input image, yielded the most promising results. However, there remain opportunities for further improvements in the architecture. One potential avenue for future research involves the design of a trainable self-attention module specifically tailored for generating masks. We feel this approach has the potential to enhance the discriminative power of the attention mask, leading to improved performance of the VPR network. Additionally, exploring alternative semantic segmentation models or incorporating other contextual information could further enhance the capabilities of the VPR network. Overall, our study provides valuable insights into the integration of attention mechanisms for visual place recognition. Alternatively, it may prove beneficial to explore the inclusion of a synthetic dataset like IDDA [12]. Such a dataset includes variations of scenes taken from identical positions and orientations but with varying time, weather, or dynamic objects (i.e. cars and people). This may help reduce biases in which certain places are highly correlated with the presence of non-static objects in the scene.

## References


[1] M. Leyva-Vallina, N. Strisciuglio, and N. Petkov, "Data-efficient large scale place recognition with graded similarity supervision," *CVPR*, 2023.

[2] V. Paolicelli, A. Tavera, C. Masone, G. Berton, and B. Caputo, "Learning semantics for visual place recognition through multi-scale attention," in *Image Analysis and Processing–ICIAP 2022: 21st International Conference, Lecce, Italy, May 23–27, 2022, Proceedings, Part II*. Springer, 2022, pp. 454–466.

[3] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," 02 2018.

[4] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017.

[5] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," 2016.

[6] G. Peng, Y. Yue, J. Zhang, Z. Wu, X. Tang, and D. Wang, "Semantic reinforced attention learning for visual place recognition," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 415–13 422.

[7] S.-M. Choi, S.-I. Lee, J.-Y. Lee, and I. S. Kweon, "Semantic-guided de-attention with sharpened triplet marginal loss for visual place recognition," *Pattern Recognition*, vol. 141, p. 109645, 2023.

[8] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "Bam: Bottleneck attention module," *arXiv preprint arXiv:1807.06514*, 2018.
[9] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.
[10] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3156–3164.
[11] F. Warburg, S. Hauberg, M. López-Antequera, P. Gargallo, Y. Kuang, and J. Civera, "Mapillary street-level sequences: A dataset for lifelong place recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2623–2632.
[12] E. Alberti, A. Tavera, C. Masone, and B. Caputo, "Idda: A large-scale multi-domain dataset for autonomous driving," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5526–5533, 2020.