2023 2nd International Conference on Artificial Intelligence, Human-Computer Interaction and Robotics (AIHCIR) | 979-8-3503-6036-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/AIHCIR61661.2023.00022

Distance-Ranking-Based Weighted Triplet Loss for Visual Place Recognition

1st Yu Xiong
Institute of Automation, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, China
xiongyu2021@ia.ac.cn

2nd Shixiong Xu
Institute of Automation, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, China
xushixiong2020@ia.ac.cn

3rd Gaofeng Meng*
Institute of Automation, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, China
gfmeng@nlpr.ia.ac.cn

***Abstract*—In the realms of computer vision and robotics, the concept of visual place recognition holds significant prominence. Its objective revolves around equipping a model with the capability to identify, from a provided query image, the most analogous images within a database, thereby identifying the place represented by the query image. Current visual place recognition models typically rely on triplet loss functions for training, but such loss functions have limitations. Traditional triplet loss functions only classify database samples into positive and negative classes, without considering the importance ranking among samples. Some samples may be more similar to the query image and contain more useful information, thus deserving more attention during training. In tackling this problem, our approach introduces a novel loss function termed the distance-ranking-based weighted triplet loss. This unique loss function assigns weightage to triplets by evaluating the spatial gap separating positive instances and the queried image, thereby intensifying the emphasis on pivotal samples. Within the framework of place recognition tasks utilizing the NetVLAD pipeline, our method achieves approximately a 1% improvement in both the Recall@1 and Recall@5 compared to traditional triplet loss function.**

Keywords-visual place recognition; triplet loss; vector of locally aggregated descriptors (VLAD)

I. Introduction

The relentless progress in deep learning and computer vision technologies has immensely driven the evolution of visual place recognition [1]-[3]. This area of study delves into discerning the exact locale depicted in an image by harnessing the visual cues embedded within it. Its utility permeates diverse sectors, including the realm of self-driving vehicles [4], virtual environments, and robotic localization systems [5].

Visual place recognition involves a matching challenge, where the primary objective is to identify the optimal pairing between a query image and images within a database. This matching process typically occurs within a feature space. The crux of this endeavor lies in developing sturdy feature representations for individual locations, ensuring their adaptability to diverse contextual shifts like changes in lighting, seasons, and viewing perspectives.

Lately, studies within the domain of visual place recognition [6]-[12] have predominantly concentrated on the extraction, selection, and amalgamation of resilient features. Additionally, there has been an emphasis on training using triplet loss [14] to minimize the distance in feature space between positive samples while increasing the distance between negative samples. This approach has yielded significant advancements in improving place recognition performance. However, limited attention has been given to the limitations of the triplet loss function in existing research.

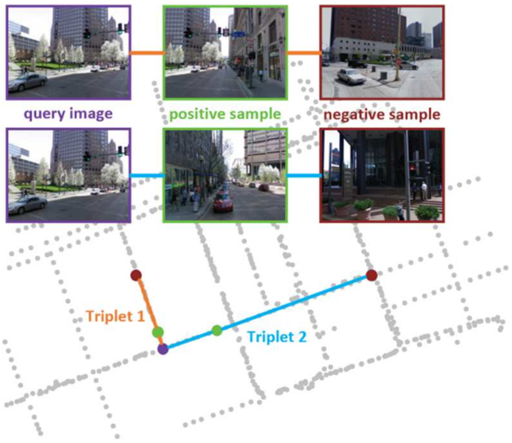

Figure 1. Illustration of triplet importance.

The limitation of triplet loss lies in its insufficient consideration of the distance ranking information between different locations. This approach merely classifies samples as positive or negative, neglecting the variations in distances among positive instances and among negative instances. In actuality, the hierarchical importance of distance rankings among distinct locations holds differing degrees of relevance for query images, and this distinction often goes unnoticed.

For example, in Fig. 1, we can observe two triplets connected by orange and blue lines respectively. A triplet is formed, including an image marked in purple as the query, another in green as a positive sample, and a third in red representing a negative sample. Examining the triplet connected by the orange line, it's evident that the positive sample image exhibits both a shorter distance from the query image and a greater degree of visual similarity. Therefore, the orange triplet carries more useful visual information and holds a higher importance ranking compared to the blue triplet. Consequently,

979-8-3503-6036-3/23/$31.00 ©2023 IEEE
DOI 10.1109/AIHCIR61661.2023.00022

during model training, more focus on these orange triplets can improve location representation learning.

To address the aforementioned issue, we propose a distance-ranking-based weighted triplet loss approach. This method fully takes into account the permutation order among different positive samples. For positive samples ranked higher in the order, we assign them higher weights, emphasizing their importance in representing the location for query images. Precisely, we allocate distinct weights to triplets contingent upon the spatial separation of the positive sample from the query image within each triplet, giving higher weights to triplets containing closer positive samples. This strategy ensures that during the gradient update process, positive samples that are closer to the query image have a greater influence. Finally, in the visual place recognition task based on the NetVLAD pipeline, our approach has shown an improvement of approximately 1% in both the Recall@1 and Recall@5 compared to the traditional triplet loss function.

II. Related Work

A. Traditional visual place recognition methods

Traditional visual place recognition methods primarily use manually designed features to represent images and then employ matching algorithms to compare image similarity. For example, the Scale-Invariant Feature Transform (SIFT) [15] method captures keypoint features that remain invariant despite changes in scale and rotation. This is achieved through a series of procedures, including the detection of extrema in scale-space, pinpointing keypoints, determining orientation, and generating descriptors. The Bag-of-words [16] approach extracts and quantizes local features in images into visual words and constructs image feature vectors by statistically analyzing the distribution of these words. Fisher vectors [17] generate high-dimensional vectors for image representation by computing the gradients of image features with respect to their Gaussian mixture models.

These features are often based on human visual perception and understanding of images, making them interpretable. However, they require manual feature design and parameter tuning, and their performance is limited by human expertise and knowledge, unable to fully exploit image information.

B. Deep learning-based visual place recognition methods

In recent times, there has been remarkable progress in the realm of visual place recognition, primarily attributed to the advancements in deep learning techniques. These methods, trained end-to-end, can automatically learn abstract and efficient feature representations and have found wide applications in visual place recognition and other fields.

Initially, methods using deep learning features primarily flattened the feature maps of Convolutional Neural Network (CNN) into feature vectors [18]-[20], which resulted in issues of high dimensionality and insufficient generalization. Subsequently, a series of new image representation methods emerged, involving operations such as aggregation and encoding of CNN feature maps to extract more distinctive and compact global feature representations. For instance, the Generalized Mean Pooling (GeM) method [21] uses generalized mean to aggregate feature maps from convolutional layers. The NetVLAD method [12] maps image features to a visual vocabulary space and applies VLAD [23] encoding to generate fixed-length feature vectors. Later on, visual place recognition further improved its performance by introducing image pyramids and attention mechanisms to more effectively utilize multi-layer CNN features and capture discriminative visual information. For example, the Deep Embedding Local Features (DELF) method [8] employs an attention mechanism to select local features, excluding irrelevant parts for the task, thereby enhancing the representation of global features. The Spatial Pyramid Encoding Vector of Locally Aggregated Descriptors (SPE-VLAD) method [9] utilizes spatial pyramid pooling to capture feature information at different resolutions and introduces a weighted triplet loss function to constrain feature distribution.

Given that these methods encode images into coarse-grained global descriptors, they have inherent limitations in their representation capabilities. Recent approaches have introduced local descriptors to further enhance performance. For example, Patch-NetVLAD [13] divides images into multiple overlapping local regions, calculates local features for each region, and performs cross-matching and geometric verification to obtain similarity scores, then reorders the top k results from NetVLAD. TransVPR [22] integrates global and local descriptors by incorporating attention masks at different levels of a Transformer [24] to perform matching. It is worth noting that there is limited research exploring this from the perspective of the loss function.

III. Methodology

Our method relies on a seamlessly integrated deep convolutional neural network structure, represented in Fig. 2. Within this architecture, triplets are initially constructed, and subsequent processes involve feature extraction through CNN, feature aggregation utilizing NetVLAD, and the computation of the triplet loss function. Subsequently, backpropagation is performed to update model parameters. However, considering the limitations of the original triplet loss, we introduce distance-based ranking information to improve the loss function, enhancing the model's performance. In the upcoming sections, a comprehensive insight into our visual place recognition framework will be presented.

A. Visual Place Recognition Framework

Due to the need to make appropriate choices between different levels of location ranges, conducting location classification tasks involves the challenge of subdividing location ranges. If locations are divided too finely, it will lead to numerous categories, making effective classification difficult. Conversely, if location ranges are divided too broadly, it will reduce the number of categories, resulting in less accurate classification. Furthermore, the boundaries between different categories are often unclear, especially when locations are situated

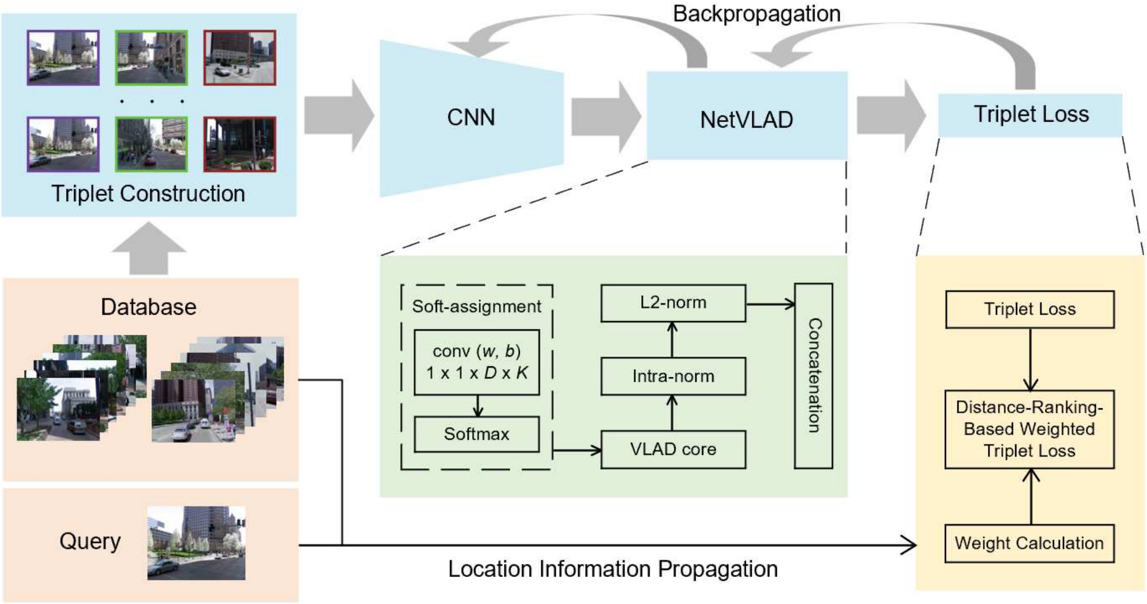

Figure 2. Overall network framework of our method.

in border areas [25], making it difficult to determine which category they should belong to.

Therefore, in visual place recognition tasks at the city level, a common approach is to treat it as a matching task. The core of this task is to construct an image database containing city images, with each image accompanied by location information such as latitude and longitude or UTM coordinates. Additionally, each location typically has multiple images covering different angles and perspectives to ensure comprehensive coverage of information. Once such a database is constructed, when there is a query image, the most similar image can be found by matching it with all the images in the database. Subsequently, the data corresponding to the image with the closest match can be harnessed to symbolize the location depicted in the query image. This methodology accomplishes the task of visual place recognition.

During image matching assignments, features from images are commonly derived utilizing neural network methodologies, followed by similarity calculations in the feature space to find the most matching image. This process requires overcoming some challenges, especially for images from the same location but with variations in perspective, lighting, and other factors. Ensuring the stability of extracted features against alterations in appearance is imperative, preventing them from being affected by surface modifications. Additionally, it is essential to ensure that images from different locations have sufficient separation in the feature space to effectively distinguish them. To impart this capability to neural networks, researchers have employed various techniques such as contrastive learning [26] and embedding learning [27].

At the heart of triplet loss lies the concept of forming numerous image triplets. Each triplet comprises a queried image, a positively selected sample image, and a negatively chosen sample image. Within these triplets, the positive sample image originates from an identical location, contrasting with the negative sample image sourced from a distinct locale. Throughout training, the primary goal is to diminish the disparity between the feature representations of the query image and the positive sample, simultaneously amplifying the gap between the feature representations of the query image and the negative sample.

After the neural network undergoes training with the triplet loss, the necessity for constructing triplets is eliminated during the inference stage. During this phase, inputting both the query image and the database images directly into the trained neural network generates their unique feature representations. Evaluating the resemblances between these features enables the identification of the most relevant image, pinpointing the location represented by the query image.

B. Original Triplet Loss

The central goal in visual place recognition is to derive distinct and resilient features from location images, intending to refine the precision in pairing query images with those stored in a database. Triplet loss is a highly effective method for achieving this [28]-[29]. The mechanism of this loss function brings locations with similar features in the embedding space closer together while pushing dissimilar locations farther apart, significantly improving the model's ability to learn representations of locations. Consequently, the triplet loss function has been extensively utilized in the domain concerning recognition of visual locations.

Within the scope of tasks related to visual place recognition, we commonly encounter a collection of query images $\{q_i, i \in [1, n]\}$ (n indicates the count of query images) and an image database Db . For constructing meaningful training triplets, it's vital to guarantee the presence of a query image, denoted as q_i, along with a corresponding positive sample image, p_i, and a negative sample image, n_i . These positive and negative samples, sourced from the database Db, need to meet the following criteria: for each triplet $\{q_i, p_i, n_i\}$, the corresponding location information satisfies

$$|y_q - y_p| < |y_q - y_n| \quad (1)$$

here y_q, y_p, and y_n denote the actual location coordinates of the query image q_i, positive sample image p_i, and negative sample image n_i, correspondingly. Therefore, the triplet set we ultimately construct can be represented as

$$T = \{(q_i, p_i, n_i) \,|\, |y_q - y_p| < |y_q - y_n|\} \quad (2)$$

During the subsequent model training, we input the images from the triplets into the model and extract corresponding feature representations, denoted as f_q, f_p, f_n. We then employ the following form of triplet loss function [14] for training:

$$L = \sum max(0,\ d(f_q, f_p) - d(f_q, f_n) + m) \quad (3)$$

Here, m is a pre-defined constant representing the margin of the loss. The function $d(a, b)$ signifies the Euclidean distance calculated between a and b.

The triplet loss function excels in learning location representations, but it is important to note that it primarily divides samples into positive and negative samples without sufficiently considering ranking or relative importance information within samples. In reality, some positive samples may be more similar to the query image or closer to it in real space, and these positive sample images contain more relevant information for the query image. This is crucial for effective learning of location representations and should, therefore, receive greater attention. However, the triplet loss does not explicitly consider this relative ranking information. Consequently, this limitation to some extent constrains the further development of visual place recognition technology.

C. Distance-Ranking-Based Weighted Triplet Loss

Triplet loss typically relies on a fixed margin value to increase the difference between positive and negative samples concerning the query image. However, it lacks a mechanism to consider that triplets selected during training may have varying levels of importance for feature learning. In practical training, the chosen triplets may contribute differently to feature learning. Nevertheless, when using traditional triplet loss, it is challenging to effectively adjust for these varying contributions.

To tackle this concern, we propose a non-uniform weighting function, denoted as $w(f_q, f_p, f_n)$, to measure the importance of each triplet. This weighting function can be defined as follows:

$$w(f_q, f_p, f_n) = \frac{1+\epsilon}{\frac{|y_q - y_p|}{\sigma} + \epsilon} - 1 \quad (4)$$

where ϵ is a minute constant employed to avoid errors resulting from division by zero. y_q and y_p represent the positional coordinates of the query image and the positive sample image, $|\cdot|$ denotes the Euclidean distance, and σ is the normalization parameter for the Euclidean distance.

By introducing this non-uniform weighting function, we can better assess the importance of each triplet, ensuring that samples with varying contributions are appropriately updated during training. This approach allows for a finer-grained control of the impact of triplets, improving the effectiveness of feature learning to better match real-world scenarios.

When attempting to directly apply weights to the triplet loss, we observed that the convergence process was not sufficiently stable. We suspected that this might be related to the chosen margin value, as the setting of the margin affects the severity of the loss. Furthermore, this margin-based approach also weakens the ordering information among triplets. Hence, we contemplated simplifying the loss function by excluding the margin parameter and concentrating on diminishing the distance, denoted as $d(f_q, f_p)$, between the query image and the positive sample image. Simultaneously, we aimed to expand the distance, represented as $d(f_q, f_n)$, between the query image and the negative sample image.

To measure these two distances, we introduced a softmax function [30] to normalize them into d_p and d_n.

$$d_p = \frac{\exp(d(f_q, f_p))}{\exp(d(f_q, f_p)) + \exp(d(f_q, f_n))} \quad (5)$$

$$d_n = \frac{\exp(d(f_q, f_n))}{\exp(d(f_q, f_p)) + \exp(d(f_q, f_n))} \quad (6)$$

Our objective is to narrow down the distance d_p, making it nearly 0 post-softmax transformation. Simultaneously, we aim to widen the distance d_n, moving it closer to 1. To achieve this goal, we employed a loss function in the form of cross-entropy (CE):

$$L_{CE} = -t_p \log(d_p) - t_n \log(d_n) \quad (7)$$

since t_p and t_n are respectively 0 and 1, this loss function can be simplified to:

$$L_{CE} = -\log(d_n) \quad (8)$$

Finally, we multiply the obtained loss function by the weighting function $w(f_q, f_p, f_n)$ to obtain the final loss function:

$$L = \sum -w(f_q, f_p, f_n) \cdot \log(d_n) \quad (9)$$

Crucially, even with the utilization of cross-entropy loss, the presence of the query image, positive sample, and negative sample within the loss function persists, retaining its triplet loss nature. This enhanced structure of the loss function meticulously considers the intricate interconnections between the query image and both the positive and negative samples.

IV. Experimental Results

In this chapter, we will examine experimental details, including the specific implementation, dataset description, evaluation metric explanations, and conduct a comprehensive analysis of experimental results, both quantitatively and qualitatively. By exploring these perspectives, our aim is to comprehend and evaluate the effectiveness of our approach in tasks related to recognizing visual locations. These analyses and results will shed light on the superiority and efficacy of our method.

A. Implementation

In our experimental setup, our model predominantly comprises three core elements: feature extraction, feature amalgamation, and a separate stage that pertains to the loss function. In the context of the feature extraction phase, we utilized VGG16 [31] as the backbone network. For feature aggregation, we explored three different aggregation methods: NetVLAD [12], max [32], and mean [33]. For the NetVLAD aggregation method, we configured the number of clusters to be 64. Regarding the loss functions, we considered four different loss functions. Firstly, the original triplet loss with a margin parameter set to 0.1, as shown in Equation 3. The second one is the original triplet weighted loss, where we directly multiplied the weighting function $\mathrm{w}(\mathrm{f}_\mathrm{q}, \mathrm{f}_\mathrm{p}, \mathrm{f}_\mathrm{n})$ (Equation 4) with the original triplet loss (Equation 3). The third is the CE (cross-entropy) loss (Equation 8), and the fourth is our proposed distance-ranking-based weighted triplet loss (Equation 9). In the weighting function $\mathrm{w}(\mathrm{f}_\mathrm{q}, \mathrm{f}_\mathrm{p}, \mathrm{f}_\mathrm{n})$, we set ϵ to 0.1 and σ to 800.

During the training process, we configured the batch size as 4, meaning each batch contained 4 triplets, totaling 12 images. Training was performed over 30 epochs, employing the stochastic gradient descent (SGD) optimizer with a learning rate set at 0.0001. Additionally, momentum was set to 0.9, and weight decay stood at 0.001. A StepLR scheduler was utilized to facilitate the model's convergence throughout the training process, where the learning rate was diminished by a factor of 0.5 after every 5 epochs. Additionally, we utilized 8 threads for training. We also set the random seed to 123 to ensure the reproducibility of the experiments. In the testing phase, we opted for the model parameters exhibiting superior performance on the validation set. These parameters were then utilized to assess the model's capabilities on the test set.

B. Dataset

Our research draws upon data from the Pittsburgh30k dataset [12], which is a large dataset widely used in visual localization and navigation research. It was created by researchers from Carnegie Mellon University and the University of Pittsburgh. This dataset comprises high-resolution images from the downtown area of Pittsburgh, covering various locations within the city, including business districts, parks, schools, restaurants, and more. Each location is represented by 24 images, which capture 12 different yaws and two pitches.

TABLE I. Dataset Quantity Distribution Table

Dataset	Query	Database
Training	7416	10000
Validation	7608	10000
Testing	6816	10000

Following the convention of the NetVLAD method [12], we partitioned the Pittsburgh30k dataset into three separate segments: a training subset, a validation subset, and a test subset. Each part contains a specific number of query images and database images, and these three parts do not overlap geographically. The specific data quantities for each set are shown in Table I.

C. Evaluation Metric

In line with typical practices, our chosen evaluation metric for the visual place recognition task is Recall@N. Here, Recall signifies the ratio of accurately matched query images to the overall count of query images. For each query image, if there is at least one true positive within the top N retrieval results, it is considered a successful match. Therefore, Recall@N can be used to measure the model's performance at different values of N, indicating whether the model can capture genuine location information within the top N results, thus providing a comprehensive assessment of the model's performance. Typically, we focus on Recall@1 and Recall@5.

D. Quantitative Results

Table II presents the experimental results for four different loss function settings. These four settings are kept identical in all aspects except for the loss function used. They all utilize VGG16 for feature extraction and employ NetVLAD for feature aggregation. Specifically, T-loss represents the original triplet loss, which serves as the baseline method. W-T-Loss denotes the weighted treatment of the original triplet loss based on the importance of triplets. CE-Loss uses the form of cross-entropy (CE) loss to minimize the distance between the query image and the positive sample while maximizing the distance from negative samples. Finally, DW-T-Loss represents our proposed distance-ranking-based weighted triplet loss.

Comparing T-loss and W-T-Loss results reveals a slight performance drop when directly weighting the original triplets. This phenomenon may be attributed to the constraint imposed by the margin parameter during the weighting process, causing the loss of some triplets to be directly set to zero and limiting the effectiveness of weighting. Consequently, we decided to eliminate the margin parameter to remove this limitation and adopted the cross-entropy (CE) loss to supervise model training. In doing so, we aimed to bring the query image as close as possible to the positive sample while pushing it further away from the negative samples. However, compared to T-Loss, the adoption of CE-Loss clearly resulted in a performance drop. This is because CE loss treats all triplets equally, and when we attempt to push the negative sample distance infinitely, it negatively impacts other triplets. To address this issue, we introduced a weighting strategy on top of the CE loss to better control the update magnitude of each triplet. Through this approach, we reduced the influence of each triplet on others, giving more weight to important triplets and less to unimportant ones. This improvement achieved approximately 1% improvement in both Recall@1 and Recall@5 performance metrics, indicating the effectiveness of the DW-T-Loss loss function.

TABLE II. THE RECALL RESULTS FOR VARIOUS LOSS SETTINGS

Loss	Recall@1	Recall@5	Recall@10	Recall@20
T-Loss	0.8121	0.9098	0.9359	0.9555
W-T-Loss	0.8110	0.9074	0.9331	0.9550
CE-Loss	0.7974	0.9026	0.9312	0.9523
DW-T-Loss	**0.8264**	**0.9139**	**0.9372**	**0.9575**

TABLE III. THE MAXIMUM AND AVERAGE AGGREGATION RESULTS FOR ORIGINAL TRIPLET LOSS AND DISTANCE-RANKING-BASED WEIGHTED TRIPLET LOSS

Method	Recall@1	Recall@5	Recall@10	Recall@20
Max	0.6381	0.8159	0.8697	0.9124
Max-W	0.6696	0.8283	0.8801	0.9181
Mean	0.5979	0.8049	0.8705	0.9178
Mean-W	0.5816	0.7861	0.8526	0.9042

To validate the effectiveness of our proposed distance-ranking-based weighted triplet loss for different feature aggregation methods, additional experiments were carried out utilizing both max aggregation and mean aggregation techniques. The outcomes of these experiments can be found in Table III. In this tabulated data, Max-W and Mean-W represent the test results obtained after training with the distance-ranking-based weighted triplet loss. For the Max aggregation method, using our proposed loss function resulted in an improvement of approximately 3% in Recall@1 and approximately 1% in Recall@5, demonstrating significant effectiveness in this scenario. However, for the Mean aggregation method, the performance declined. We speculate that this phenomenon may be due to the nature of global average aggregation itself. Many interference elements exist in images, and we are more concerned with the primary features rather than all features. Therefore, averaging all features may weaken the effectiveness of image representation, and adopting a weighting strategy may not only improve performance but also potentially lead to a performance decline.

E. Qualitative Results

We illustrated the outcomes from our proposed technique and the initial approach to image matching, depicted in Fig. 3. In this visual representation, the initial column exhibits the query images, the second column illustrates the top-1 matches generated by the model trained using the distance-ranking-based weighted triplet loss, and the third column showcases the top-1 matches produced by the model trained with the original triplet loss. Correct matches are denoted by green boxes, whereas incorrect matches are marked with red boxes.

Examining the outcomes from the initial row, it becomes apparent that our innovative approach empowers the model to concentrate more efficiently on essential discriminative details within the images (such as buildings), while disregarding extraneous and irrelevant elements (such as roads and bridges). In the second row, our method demonstrates adaptability to certain variations in appearance, such as changes in storefront design. The results in the third row illustrate our method's ability to handle variations in lighting, successfully matching images of locations with different color tones. These visuals demonstrate our approach's effectiveness.

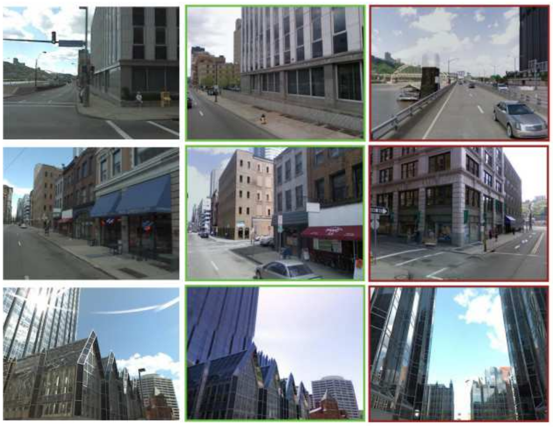

Figure 3. Qualitative results figure.

V. CONCLUSION AND FUTURE WORK

The objective of this research is to elevate the training methodology employed in visual place recognition tasks, with the aim of enhancing the overall performance of the model. Traditional triplet loss functions categorize samples into positive and negative classes during training, neglecting the importance and ranking information among samples. To overcome this limitation, we introduce the distance-ranking-based weighted triplet loss, which assigns weights to triplets by considering the distance between positive samples and query images, thus focusing more on important samples. In the visual place recognition task using the NetVLAD pipeline, our approach achieves approximately a 1% improvement in performance on both Recall@1 and Recall@5 metrics. This research provides new insights and directions for enhancing visual place recognition methods.

In the future, we can further research and enhance the distance-ranking-based weighted triplet loss function to further improve the performance of visual place recognition models. We can explore different weighting strategies and loss function variants to find more effective training methods. Additionally, we can investigate the application of this method in other related fields, such as image retrieval and geographic information systems, to expand its scope of application. Furthermore, future research can focus on handling larger and more diverse location databases to meet the demands of practical applications. Moreover, we can consider incorporating multimodal information, such as textual descriptions or voice commands, to further improve the accuracy and robustness of place recognition.

REFERENCES

[1] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, et al. (2015) "Visual place recognition: A survey." IEEE Transactions on Robotics, 32: 1-19.

[2] X. Zhang, L. Wang, and Y. Su. (2021) "Visual place recognition: A survey from deep learning perspective." Pattern Recognition, 113: 107760.

[3] C. Masone and B. Caputo. (2021) "A survey on deep visual place recognition." IEEE Access, 9: 19516-19547.

[4] G. Bresson, Z. Alsayed, L. Yu and S. Glaser. (2017) "Simultaneous localization and mapping: A survey of current trends in autonomous driving." IEEE Transactions on Intelligent Vehicles, 2: 194-220.

[5] E. Stumm, C. Mei, and S. Lacroix, 2013. "Probabilistic place recognition with covisibility maps." In: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. Tokyo. 4158-4163.

[6] H. Jin Kim, E. Dunn, and J.-M. Frahm, 2017. "Learned contextual feature reweighting for image geo-localization." In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu. 2136-2145.

[7] Y. Zhu, J. Wang, L. Xie, and L. Zheng, 2018. "Attention-based pyramid aggregation network for visual place recognition." In: Proceedings of the 26th ACM international conference on Multimedia. Seoul. 99-107.

[8] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, 2017. "Large-scale image retrieval with attentive deep local features." In: Proceedings of the IEEE International Conference on Computer Vision. Venice. 3456-3465.

[9] J. Yu, C. Zhu, J. Zhang, Q. Huang, and D. Tao. (2019) "Spatial pyramid-enhanced NetVLAD with weighted triplet loss for place recognition." IEEE Transactions on Neural Networks and Learning Systems, 31: 661-674.

[10] G. Peng, Y. Yue, J. Zhang, Z. Wu, X. Tang, and D. Wang, 2021. "Semantic reinforced attention learning for visual place recognition." In: 2021 IEEE International Conference on Robotics and Automation (ICRA). Xi'an. 13415-13422.

[11] G. Peng, J. Zhang, H. Li, and D. Wang, 2021. "Attentional pyramid pooling of salient visual residuals for place recognition." In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal. 885-894.

[12] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, 2016. "NetVLAD: CNN architecture for weakly supervised place recognition." In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas. 5297-5307.

[13] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, 2021. "Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Online. 14141-14152.

[14] F. Schroff, D. Kalenichenko, and J. Philbin, 2015. "Facenet: A unified embedding for face recognition and clustering." In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston. 815-823.

[15] D. G. Lowe. (2004) "Distinctive image features from scale-invariant keypoints." International Journal of Computer Vision, 60: 91-110.

[16] J. Sivic and A. Zisserman, 2003. "Video Google: A text retrieval approach to object matching in videos." In: Proceedings Ninth IEEE International Conference on Computer Vision. Nice. 1470-1470.

[17] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier, 2010. "Large-scale image retrieval with compressed fisher vectors." In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco. 3384-3391.

[18] Z. Chen, O. Lam, A. Jacobson, and M. Milford. (2014) "Convolutional neural network-based place recognition." arXiv preprint arXiv:1411.1509.

[19] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, 2015. "On the performance of convnet features for place recognition." In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Hamburg. 4297-4304.

[20] T. Naseer, M. Ruhnke, C. Stachniss, L. Spinello, and W. Burgard, 2015. "Robust visual SLAM across seasons." In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Hamburg. 2529-2535.

[21] F. Radenović, G. Tolias, and O. Chum. (2018) "Fine-tuning CNN image retrieval with no human annotation." IEEE Transactions on Pattern Analysis and Machine Intelligence, 41: 1655-1668.

[22] R. Wang, Y. Shen, W. Zuo, S. Zhou, and N. Zheng, 2022. "TransVPR: Transformer-based place recognition with multi-level attention aggregation." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans. 13648-13657.

[23] H. Jégou, M. Douze, C. Schmid, and P. Pérez, 2010. "Aggregating local descriptors into a compact image representation." In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco. 3304-3311.

[24] A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, et al. (2021) "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929.

[25] G. Berton, C. Masone, and B. Caputo, 2022. "Rethinking Visual Geo-localization for Large-Scale Applications." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans. 4878-4888.

[26] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, 2020. "Momentum contrast for unsupervised visual representation learning." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle. 9729-9738.

[27] H. Chen, B. Perozzi, R. Al-Rfou and S. Skiena. (2018) "A tutorial on network embeddings." arXiv preprint arXiv:1808.02590.

[28] L. Wang, Y. Li, and S. Lazebnik, 2016. "Learning deep structure-preserving image-text embeddings." In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas. 5005-5013.

[29] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, et al, 2014. "Learning fine-grained image similarity with deep ranking." In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus. 1386-1393.

[30] E. Simo-Serra and H. Ishikawa, 2016. "Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction." In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas. 298-307.

[31] K. Simonyan and A. Zisserman. (2014) "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556.

[32] H. Azizpour, A. Sharif Razavian, J. Sullivan, A. Maki, and S. Carlsson, 2015. "From generic to specific deep representations for visual recognition." In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Boston. 36-45.

[33] A. Babenko and V. Lempitsky, 2015. "Aggregating local deep features for image retrieval." In: Proceedings of the IEEE International Conference on Computer Vision. Santiago. 1269-1277.