Feature-Level Knowledge Distillation for Place Recognition based on Soft-Hard Labels Teaching Paradigm

Zhenyu Li, *Member, IEEE*, Pengjie Xu, Zhenbiao Dong, Ruirui Zhang, and Zhaojun Deng

***Abstract*—The motivation of visual place recognition (VPR) is to enable robots to identify and localize specific places within an environment using visual cues, facilitating navigation, mapping, and context-aware applications. On the other hand, deeper networks impose an extra computing strain on robots and greatly impede the development of real-time robot applications. Most groundbreaking studies either focus on learning from only one teacher in their distilled approach to learning, ignoring the possibility of students learning from multiple teachers, or failing to reveal teachers place varying importance on specific examples. To cope with the above issues, we propose a novel adaptive soft-hard label teaching feature-level knowledge distillation learning framework, namely ASHT-KD, for all-day mobile robot VPR tasks. This framework learns a compact and quick all-day place recognizer through knowledge transfer from several teachers to a limited number of students. Specifically, depending on the complexity of the environments, teachers can impart knowledge to two types of students in two teaching modes: soft-label teaching and hard-label teaching, which corresponds to one type of student being required to learn a new and uncomplicated environment (query image), while the other type of students are forced to learn a more complex environment (database images). To balance computational memory and performance, the teacher network is designed to be a two-level sampling ViT pipeline, while the Siamese student network is constructed to be a lightweight pipeline consisting only of one-level down-sampling ViT for place matching. In addition, a cross-entropy loss network is introduced to further improve the VPR performance by strengthening the correlation of feature representations from the Siamese network. Extensive experiments demonstrate the effectiveness and superiority of ASHT-KD. The practicability of *ASHT-KD* is also verified through outdoor testing.**

***Index Terms*—mobile robot, visual place recognition, soft-hard label-based knowledge distillation, vision Transformer, and all-day place recognizer.**

I. Introduction

VISUAL place recognition is crucial for intelligent mobile robot applications in visual localization and navigation, can assist the robot in eliminating drift errors, and is usually used for SLAM loop-closure detection, surveying [1], [2], [3]. VPR aims to identify the query image in unknown regions by matching it with a set of reference images in known places to judge whether it has been previously visited using visual information. Performing a highly robust and reliable VPR is a critical component of autonomous robot navigation, as simultaneous localization and mapping (SLAM) systems rely on a closed-loop mechanism for map correction. Although the VPR problem is well-defined and widely studied, performing it reliably remains an extremely difficult task as a series of challenges must be coped with. First, seasonal changes, viewpoint changes, illumination changes, dynamic elements, or any combination of these factors, result in the places that have been visited may appear significantly different from the places first seen or recorded. On the other hand, Similarity across places can lead to perceptual confusion, especially within the same environment. However, due to the influence of the above factors, there may be low brightness, low contrast, appearance distortion, and low signal-to-noise ratio among the places visited by mobile robots, which seriously degrades the performance of existing SOTA recognizers in complex cross-domain environments. Therefore, the backbone using ideal places training loses its effectiveness in recognizing complex environmental features, leading to the failure of the mobile robot VPR task.

To maintain the excellent performance of the recognizer in cross-domain environments, a promising solution is to transfer the knowledge learned by the ideal environment recognizer to the complex environment recognizer through transfer learning with a plug-and-play strategy [4], [5], [6]. Through this operation step, the application scenarios of VPR are extended to complex cross-domain environments, and the performance can be improved compared to directly applying the model to complex cross-domain environments. Therefore, The daytime recognizer can be enhanced by transferring the knowledge learned in night scenes to itself, achieving excellent results. However, the gap in the transfer of knowledge learned in different environments is always difficult to eliminate, making it difficult for existing SOTA recognizers to improve VPR performance further. In addition, due to limited computing resources, algorithms with high computational requirements become impractical on mobile robots. The additional computational complexity of adding environmental enhancement

This work was funded by the Qing Chuang Plan by the Department of Education of Shandong Province (24240904, 24240902), the Chinese Society of Construction Machinery Young Talent Lifting Project (CCMS-YESS2023001), the Opening Foundation of Key Laboratory of Intelligent Robot (HBIR202301), the Open Project of Fujian Key Laboratory of Spatial Information Perception and Intelligent Processing (FKLSIPIP1027). *Corresponding authors: Zhenyu Li.*

Zhenyu Li and Ruirui Zhang are with the School of Mechanical Engineering, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China, and are with the Shandong Machinery Design and Research Institute, Jinan 250031, China (e-mail: lizhenyu@qlu.edu.cn,).

Pengjie Xu is with the School of Mechanical Engineering, Shanghai Jiaotong University, Shanghai 200030, China (e-mail: xupengjie194105@sjtu.edu.cn.)

Zhenbiao Dong is with the School of Mechanical Engineering, Shanghai Institute of Technology, Shanghai 201418, China (e-mail: dzb0312@126.com)

Zhaojun Deng is with the School of Mechanical Engineering, Tongji University, Shanghai 201804, China (e-mail: dengzhaojun@tongji.edu.cn)

modules hinders the development of all-day mobile robotic applications. *How to construct a knowledge transfer model without gaps or withstand small gaps within a range to achieve a robust all-day place recognizer with high performance is an urgent challenge*.

Knowledge distillation is a model compression technique that involves training a compact, lightweight model (referred to as the student model) using supervised information from a larger, high-performing model (referred to as the teacher model) to improve performance and accuracy [7], [8]. This method has seen widespread application in natural language processing (NLP) [9], object detection [10], image segmentation [11], and so on. In knowledge distillation learning, single-teacher models have some drawbacks compared to multi-teacher models. A single-teacher model offers a singular perspective, which may lead to a lack of diversity and generalization ability in the student model, making it prone to overfitting the teacher model's errors. It also heavily relies on the performance of the single teacher model, resulting in insufficient stability and robustness. In contrast, multi-teacher models combine the knowledge and strengths of multiple models, providing more comprehensive and diverse guidance, reducing individual model biases, and enhancing the student model's generalization ability and stability, thereby improving overall performance. The multi-teacher model effectively reduces the risk of overfitting and improves the adaptability of the model in different scenarios by assigning different weights to multiple teacher networks.

This paper proposes a novel adaptive soft-hard label teaching feature-level knowledge distillation learning framework (ASHT-KD) for all-day mobile robot VPR tasks. Soft-label teaching and hard-label teaching, correspond to one type of student being required to learn a new and uncomplicated environment (query image), while the other type of student is forced to learn a more complex environment (database images). We believe the proposed method is versatile and can be applied to various visual detection tasks such as object tracking [12], [13], segmentation [14], and image retrieval [15], with straightforward modifications. ASHT-KD associates each teacher with latent feature descriptions to adaptive learn feature-level teacher importance weights, leveraging these weights to acquire integrated high-level knowledge and collecting from multiple teachers through the proposed multi-group guidance strategy different levels of knowledge. In summary, our main contributions are as follows:

- A novel adaptive soft-hard label teaching feature-level knowledge distillation learning framework (ASHT-KD) has been developed to accomplish robust all-day VPR with a lightweight Siamese student model.
- A new multi-teaching pipeline is constructed to facilitate mutual the complementation of strengths among teachers, achieving comprehensive guidance to the students and superior performance in nighttime robot VPR.
- Extensive experiments on several public datasets are conducted to verify the effectiveness of our ASHT-KD, as well as an all-day place dataset is built for the real-world test.

II. Related Work

A. Robot Visual Place Recognition

VPR has been widely used in mobile robot localization and navigation (SLAM). As robots continue to develop toward intelligence, the ViT network-based method has gradually become the mainstream framework to achieve high-performance VPR for mobile robots due to its superior success rate and accuracy. Y. Wang *et al.* [16] proposed a novel transformer-based VPR architecture that combines low-level local details, spatial context, and high-level semantic information to improve robustness to appearance and viewpoint changes. S. Zhu *et al.* [17] proposed a unified VPR framework that uses a new transformer model, namely R^2Former, to address retrieval and reranking. R. Wang *et al.* [18] proposed a novel transformer-based VPR model called TransVPR, which can adaptively extract robust image representations from different regions of the image. A. Ali-Bey *et al.* [19] proposed MixVPR, a new holistic feature aggregation technique for VPR tasks, which takes feature maps from the pre-trained backbone as a set of global features. *Although these place recognizers perform well in better illumination conditions, these VPR methods struggle to efficiently perform robust VPR tasks in the low light of the night due to the huge gap between daytime and nighttime images.*

B. Knowledge Distillation

The output of the teacher model is used as a "soft target" to train the student model during the knowledge distillation process. In this way, the student model can learn the generalization ability of the teacher model, rather than just fitting the training data. With the continuous development and maturity of the knowledge distillation system, knowledge distillation has been widely used in various visual tasks [20]. A. Chawla *et al.* [10] proposed the DeepInversion for object detection to enable data-free knowledge distillation of neural networks trained for object detection tasks. Y. Feng *et al.* [21] proposed a pixel-level similarity distillation, namely PSD, that utilizes residual attention maps to capture more detailed spatial dependencies across multiple layers. W. Chen *et al.* [22] extracted semantic relevance knowledge from the extracted representations of new data for image retrieval to regularize parameter updates using knowledge distillation based on the teacher-student framework. S. Lin *et al.* [23] proposed a new knowledge sublimation method for one-to-all space matching. Specifically, the method allows each pixel of the teacher feature to be extracted to all spatial locations of the student feature as its similarity is generated by the target-aware transformer. Shen *et al*. [24] proposed a refined Siamese tracking framework to train a compact and precise tracker (student) that acquires essential knowledge from a larger Siamese tracker (teacher) through a teacher-student knowledge distillation model. *However, these methods only involve a teacher model using a fixed distillation method in the training process, which limits the potential for multi-dimensional development of the student model. This limitation is particularly evident in generalized applications in complex*

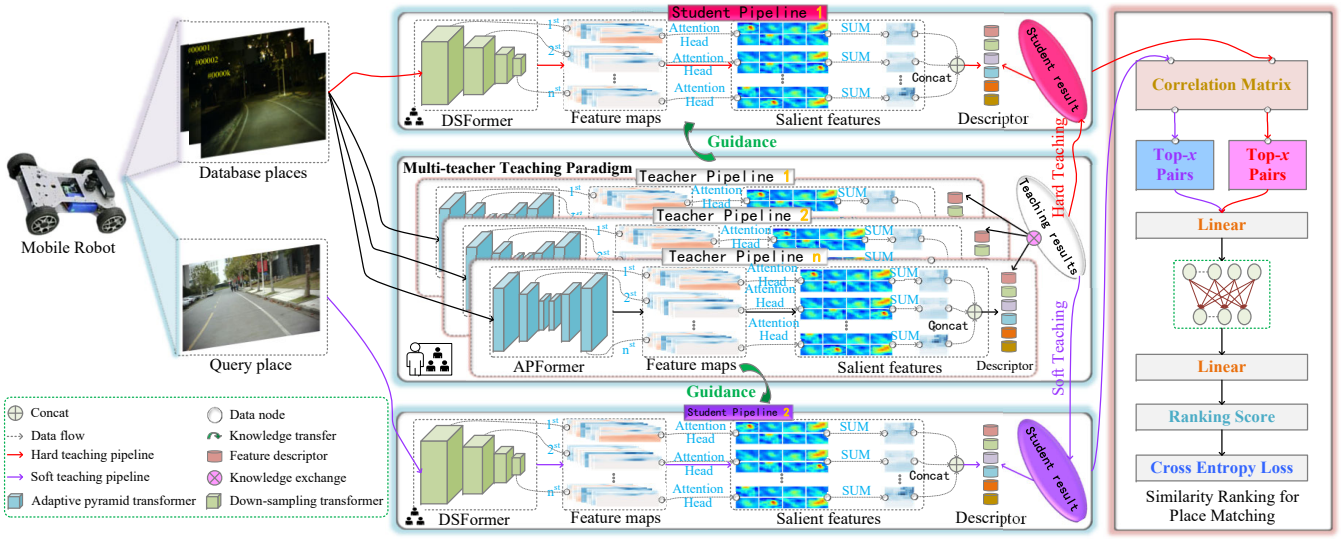


Fig. 1. An overview of the proposed knowledge distillation framework for mobile robot VPR. ASHT-KD contains several advanced teacher models (APFormer) and two student models (DSFormer). Two student models learn from multiple teacher models in a soft label teaching and a hard label teaching manner, respectively, and have a specific focus. During the training phase, multiple teacher models exchange teaching experiences in each frame and pass on the best learning methods to the two student models. In the test phase, the learned student model can directly retrieve the visited database place from the complex environment image.

VPR scenarios for mobile robots at night. Shen et al. [24] proposed a refined Siamese tracking framework to train a compact and precise tracker (student) that acquires essential knowledge from a larger Siamese tracker (teacher) through a teacher-student knowledge distillation model. This model serves as a reference for the policy-based knowledge distillation of multi-teacher teaching in our future work.

III. Methodology

This paper proposes a novel adaptive soft-hard label teaching feature-level knowledge distillation learning framework (ASHT-KD) for all-day mobile robot VPR tasks. The framework consists of multiple teacher models and two student models, as illustrated in Fig. 1. In this knowledge distillation framework, the knowledge transferring is divided into two aspects, i.e., 1) the knowledge transfer from the teacher to the students, and 2) the knowledge transfer among the teacher models.

A. Motivation and Overview

We propose a new multi-teacher collaborative teaching model that is an ensemble learning method that combines knowledge from multiple teacher models to train a student model. Soft and hard teaching are components of this method, where soft teaching uses the probability distributions output by the teacher models to guide the student model, while hard teaching uses the actual class labels for training. During the training process, the collaborative exchange of teaching experiences among a group of quantified teachers forms useful importance-aware shared knowledge at a certain layer, which is refined into the corresponding feature layer of the student network. Secondly, teachers impart knowledge to students according to the complexity of the application scenario in the form of soft label teaching and hard teaching. The VPR matching paradigm is built using a Siamese network comprising a soft label network (student A) and a hard label network (student B). This approach entails inputting the image into two sub-networks (student models) with shared weights to extract high-dimensional feature vectors. The network is trained using the cross-entropy loss function to ensure that feature vectors at the same place closer, closer while those at different places are apart. This which integrated combines leverages probabilistic information and precise classification improve enhance the accuracy and robustness of the model.

Furthermore, the design of our framework enables end-to-end training, whereby not only the teacher and student network but also the teacher's contribution to shared knowledge is learnable during the learning process.

B. Multi-Teacher Teaching Paradigm

The selection of teacher models is an important part of knowledge refinement. If there is a large ability gap between teachers and students, the teacher's knowledge may not be well transferred to the students. To restrain teacher power, we consider teacher and student models with the same architecture in our work. Teachers' knowledge from the same level will be combined to form shared knowledge, which will then be used as input by teachers at the next level, which encourages teachers to work together to learn.

Formally, given the quantification function $Q(x, b)$, $Q(W_i^j, b_i^j)$ and $Q(A_i^j, b_i^j)$ are used to quantify the i^{th} teacher, where W_i^j, A_i^j and b_i^j respectively represent the weight, activation of feature map, and bias of j^{th} feature map output layer. The shared knowledge S_j of n teachers at the level

corresponding to the j^{th} layer index is expressed as:

$$\begin{array}{c} S_j = \sum_{i=1}^{n} (\varphi_i^j) * Q(A_i^j, b_i) \\ \sum_i (\varphi_i^j) = 1, \varphi_i^j \in [0, 1] \end{array} \tag{1}$$

Where φ_i^j is the importance factor of i^{th} teacher.

To handle the constraints on the important factors of the teacher model in equation (1), the Softmax function is applied to the important factor values before they are used to calculate S_j in the implementation. End-to-end training optimizes important factors and models the weight of teachers and students at the same time.

In order to share the knowledge acquired by the multiple teacher models, the mutual-learning room of multiple teacher models is used to select the best teacher model for a specific framework to guide the student model training. Specifically, teacher model uses the attention map to determine whether the pixel regions corresponds to the target domain. For each framework, the best teacher model transfers its knowledge to other teacher models by means of another correlation. Simultaneously, important factor φ_i^j is devised to consider this factor for the teacher model:

$$\mathcal{L}_{\mathrm{ML}} = \frac{1}{n} \sum_{i=1}^{n-1} \left(\mathrm{MSE} \left(\hat{\mathcal{A}}_i^j, \mathcal{A}_i^j \right) \right) \tag{2}$$

Where $\hat{\mathcal{A}}$ and $\mathcal{A}$ represent the attention feature maps of the best teacher and student, n is the number of teacher models. It should be noted that the parameters of the best teacher model will be frozen during the backpropagation of this correlation loss.

Using the ratio of the first two maximum attention values as the persuasion (Q_F/Q_S), the best teacher model can be chosen frame by frame to assist other teacher models in learning from its strengths. Consequently, all teacher models can acquire more comprehensive knowledge compared to those that do not learn from each other, and can transmit the best knowledge to student models.

Our knowledge distillation framework is a feature-level knowledge learning and transfer process. The quantization function maps the value $x^{'} \in R$ to the quantized value $x^{'} \in \{q_1, q_2, \ldots, q_n\}$, using the quantization function Q with the exact feature dimension k. The quantified value is defined as:

$$x^{'} = Q(x, b) \tag{3}$$

This paper utilizes half-wave Gaussian quantization (HWGQ) [24] as the quantizer in our VPR framework since it is an efficient and straightforward uniform quantization approach. HWGQ first computes the optimum values q_i using uniform quantization of a unit Gaussian distribution before quantifying the weights and activations. The quantized value of x is given as the variance δ of the weights and activations:

$$x^{'} = \delta * q_i \tag{4}$$

C. Feature-Level Knowledge Distillation

Teachers' knowledge will be used to facilitate student learning. In the j^{th} layer model, let S_j^t and S_j^s be the shared feature maps of teachers and students, respectively. Let χ denote the layer index chosen for distillation based on intermediate visual features. The loss of knowledge distillation can be defined as follows:

$$L_f = \sum_{j \in \chi} \mathcal{D}(S_j^t, S_j^s) \tag{5}$$

Where $\mathcal{D}$ is the distance loss that measures the similarity between the features learned by the teacher and the student.

In the VPR task, the Siamese student models have different input objects, that is, they receive places with better illumination and low illumination, respectively, resulting in the focus of the teacher's teaching tasks being different. In our work, hard teaching is used to strengthen student learning for student models learning complex low-light environments and soft label teaching is used for student models learning high-light environments with better illumination. Comparing the two teaching paradigms, hard teaching is more difficult than soft label teaching. Therefore, The losses of knowledge distillation for soft label teaching and hard labelteaching can be defined as follows:

$$L_f^{soft} = \ell_s \cdot \sum_{j \in \chi} \mathcal{D}(S_j^t, S_j^s) \tag{6}$$

$$L_f^{hard} = \ell_h \cdot \sum_{j \in \chi} \mathcal{D}(S_j^t, \eta \cdot S_j^s) \tag{7}$$

Where η is the factor of environment complexity, ℓ_s and ℓ_h are the coefficients of soft-label and hard-label teaching teacher models.

Let Q_j^t and Q_j^s represent the attention feature maps in j^{th} layer. In this paper, attention loss is adopted to measure the similarity between teacher features and student features, which is expressed as:

$$\mathcal{D}_{AT_{Loss}} = \sum_{j \in \chi} \left\| \frac{Q_j^t}{\left\| Q_j^t \right\|_2} - \frac{Q_j^s}{\left\| Q_j^s \right\|_2} \right\|_p \tag{8}$$

Where $\|\cdot\|_p$ is the $l_2 - normalization$ function. The attention map $Q \in R^{C \times H \times W}$ has C channels, which is defined as:

$$Q = \sum_{j=1}^{C} |A_j|^p \tag{9}$$

D. Teaching-Learning Feedback Mechanism between Students and Teachers

In addition to feature-based distillation, we also utilize mutual learning [25] for feedback exchanges between teachers and students. Through mutual learning, students can give feedback on their learning status to the teacher, which encourages teachers and students to adjust their parameters at the same time to achieve an adaptive learning goal. We adopted KDCL-MinLogit, which is a simple and effective teacher-student logit integration method. In particular, this method selects the smallest logit value for each category.

The logit outputs of the cooperative teacher model and the student model are represented as z^t and z^s. $z^{t,C}$ and $z^{s,C}$ are donated as the elements of z^t and z^s that corresponds to the target dimension feature space C_d. Traditional supervised losses train the network to predict the correct labels for training

instances. To improve the generalization performance of the overall model on test instances, we use soft-hard teaching peer-to-peer networks to provide training experience in the form of soft-label teaching posterior probability p_2. To quantify the match between soft label teaching network (student 2 pipeline) and hard teaching network (student 1 pipeline) predictions p_2 and p_1, we use Kullback Leibler (KL) [26] divergence that can be expressed as:

$$D_{KL}\left(\boldsymbol{p}_2 \| \boldsymbol{p}_1\right)=\sum_{i=1}^{N} \sum_{d=1}^{D} p_2^d\left(\boldsymbol{x}'_i\right) \log \frac{p_2^d\left(\boldsymbol{x}'_i\right)}{p_1^d\left(\boldsymbol{x}'_i\right)} \tag{10}$$

Where $d \in D$ is the dimension of the feature map. Finally, Our mutual-learning loss based on KL divergence is asymmetric and hence different for the two networks. The symmetric Jensen-Shannon divergence loss can be used [27]:

$$L_{overall}=\ell_{ml} \cdot \frac{1}{2}(D_{KL}(\boldsymbol{p}_1||\boldsymbol{p}_2))+(D_{KL}(\boldsymbol{p}_1||\boldsymbol{p}_2)) \tag{11}$$

Whereℓ_{ml} is the coefficient of mutual learning.

E. Similarity Matching

After knowledge distillation, it is easily obtained the descriptors from the two student channels. Global retrieval aims to learn an embedding space in which each query image I_q is close to its corresponding database image I_q. Given a query image $\{I_q\}$ and database images $\{I_b\}$, it is defined that the global feature representations of query and database samples as F_q (student 2 pipeline), F_b (student 1 pipeline), and the global place match loss can be computed by the l_2-norm loss:

$$L_{global}=max(\|F^q-F^b\|^2) \tag{12}$$

Where $\|\cdot\|^2$ denotes squared $L2-Norm$. To obtain robust feature representation in the dark scene, we propose an adaptive feature scale coding method to learn discriminative features. Therefore, in addition to adopting global features, local features are also used for visual ranking, which is expressed as:

$$L_{local}=max(\|F_q^k-F_b^k\|^2) \tag{13}$$

Where $F_{q'}$, $F_{b'}$ are local feature representations of query and database samples, k is the k^{th} layer feature representation. Then, the reranking module combines the local features and global features of the two images to generate two-logit scores L_{cross} as the output, which is expressed as the likelihood of true or false matching:

$$L_{cross}(L_{global},L_{local})=-\ell_{cn} \cdot \sum_{\mathrm{i}=1}^{\mathrm{n}}(P_{L_{global}} log(P_{L_{local}})) \tag{14}$$

Where ℓ_{cn} is the coefficient of cross-entropy loss.

F. Adaptive Cross-weight Pyramid Transformer Backbone

To improve the effect of knowledge distillation and the computational efficiency of the overall model, in line with the principle of saving computing costs and lightweight and efficient deployment, we take the complex APFormer with up and down sampling as the teacher model, and the lightweight DSFormer with only down sampling as the student model, which is shown in Fig. 2.

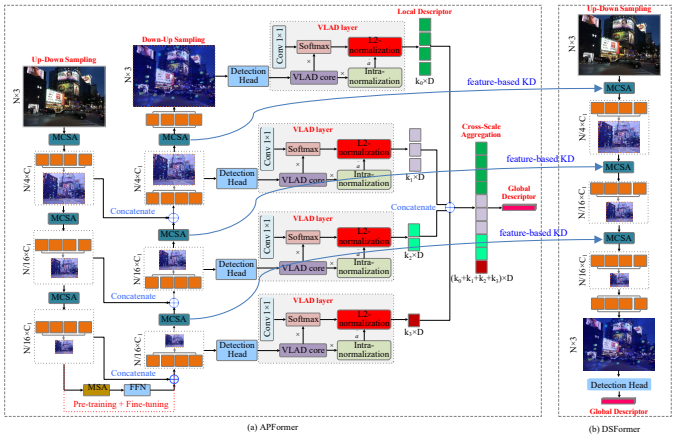


Fig. 2. Adaptive pyramid transformer model. We propose two feature extraction backbones for knowledge distillation: (a) APFormer, which is used for the teacher pipeline, and (b) DSFormer, which is used for the student pipelines.

To focus on the self-adaptive feature scale of input images, APFormer is mainly designed for automatically adjusting and balancing the proportion of intra-scale and inter-scale features in feature aggregation.

Specifically, given a query place image I_q, At the 1^{st} scale stage, $\frac{N}{4}$ feature map is obtained by the MCSA module. Similar to sequence-to-sequence learning, we adopt MCSA to project source channel features and one-hot label vectors into a dense embedding space to obtain more discriminative representations. Therefore, based on the same principle, $\frac{N}{2^4}$, $\frac{N}{2^5}$, ..., $\frac{N}{2^{w+2}}$ feature maps are obtained sequentially at the 2^{th}, 3^{th}, ..., w^{th} scale, where $w(\geq 2)$ is scale factor. Inspired by UNet, we construct fully symmetric up-sampling to encode the down-sampled feature maps of the previous stages. Therefore, by adopting the same sampling principle as down-sampling, $\frac{N}{2^{w+2}}$, ...,$\frac{N}{2^4}$, $\frac{N}{2^2}$ feature maps are obtained sequentially at the 2^{th}, 3^{th}, ..., w^{th} scale at the up-sampling stage. The above down-up sampling process can be formulated as (3) and (4):

$$\begin{aligned} f(I)_i^{down}&=MCSA(f(I)_{i-1}^{down}),\\ with f(I)_{i=0}^{down}&=I \in R^{H \times W \times C} \end{aligned} \tag{15}$$

$$\begin{aligned} f(I)_j^{up}&=MPL(Concat(f(I)_j^{up},down(f(I)_{w-j}^{down}))),\\ with f(I)_{j=0}^{up}&=MPL(FFN(f(I)_W^{down})) \end{aligned} \tag{16}$$

Where $MCSA$ denotes the multi-channel self-attention layer, FFN denotes the feed-forward neural network layer, and $Concat$ is the aggregation function.

Therefore, the teacher backbone can be described as the adaptive symmetric sampling paradigm (equation 14→15), namely APFormer and the student backbone can be described as the single-sampling paradigm (equation 14), DSFormer with an extra network layer resized feature map. As shown in the model-based attention feature heatmap shown in Fig. 3, the final best student model, namely ASHT-KD, can acquire

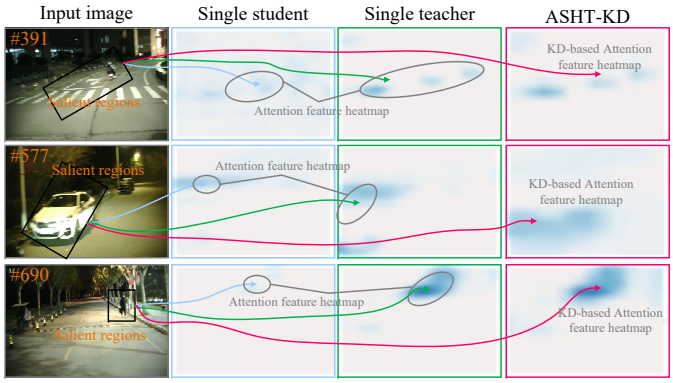


Fig. 3. Comparison of the attention feature heatmaps between ASHT-KD, single-student model, and single-teacher model.

knowledge beyond the teacher, and in some cases even performs better than the teacher.

IV. Experiments

In this section, Siamese student models are trained through the adaptive soft-hard label teaching-based knowledge distillation framework, namely ASHT-KD. ASHT-KD is compared with the existing SOTA VPR approaches, and also tested in a real environment for validation of effectiveness.

A. Implementation Details

The basic scale factor of ASHT-KD (APFormer and DSFormer) is set to $w = 128$. The input images are resized as 224×224. In the channel attention layer, we use a sliding window with the size of 4×4 to perceive the feature map of each stage. We set the batch size as 64 and selected $AdamW$ [28] as the optimizer. We initialize our model using the pretraining weight and initialize the learning rate as 0.0025 in the first 18000 iterations, and the new learning rate as 0.005 in the next 22000 iterations. To obtain better experimental results, we set the coefficients of knowledge distillation loss (including soft-label loss and hard-label loss), mutual-learning loss, and cross-entropy loss as $\ell_s = 550$ & $\ell_h = 650$, $\ell_{ml} = 0.5$, and $\ell_{ce} = 1$, respectively. The factor of environment complexity is set as $\eta = 0.6$. We conduct all training and testing processes on four NVIDIA 3060 GPUs with a total memory of 48G. The teacher model is based on up-sampling and down-sampling pyramid transformers, and the student model is based only on down-sampling transformers and uses a single layer of network to restore the size of feature maps. The parameters of the teacher model are frozen during the training phase, while the parameters of the student model are also frozen except for the VPR backbone. The mutual learning knowledge distillation framework includes n teacher models and Siamese student models.

B. Evaluation Metrics and Benchmarks

To objectively evaluate the performance of the proposed method, some publicly available evaluation metrics must be used. To calculate the percentage of correctly matched queries, we use the $Recall@N$ algorithm. Namely, $Top-1$ is the probability that the recognition result of one candidate scene is correct, and $top-5$ is the probability that the results of the predicted five candidate scenes contain at least one correct result. To evaluate the performance of the proposed ASHT-KD VPR, We use public datasets (including the KITTI_00 sequences [29], Tokyo 27/4 [30], VPRICE [31], and Nordland [32] for evaluation benchmarks:

(1) KITTI has 22 sequences containing a total of 44182 stereo images (39.2km). The dataset scene contains real images collected from urban, rural, and highway scenes. The resolution of all images is 1242×375.

(2) Tokyo 24/7 uses mobile phone cameras to collect scene that contains 76k referenced scenes and 315 query scenes.

(3) VPRICE consists of 7778 images from various outdoor environments and observation conditions. The resolution of all images is 640×480.

(4) Nordland [49] contains a 728km long cycle that is traversed 4 times. The appearance of dataset scenes varies widely, from snow-covered winter to fresh green vegetation in spring and summer to colorful foliage in autumn. All images have a resolution of 640×480.

C. Performance Comparison with SOTA Approaches

It can be seen in Fig. 4 that the proposed ASHT-KD VPR performs better than the SOTA VPR approaches, i.e., NetVLAD [33], MixVPR [19], TCL [34], TransVPR [18], Hybrid CNN-Trans [16], ETR [35], TransVLAD [36], and R^2Former [17]. To evaluate the performance of the VPR approaches more broadly, we compare methods CNN-based, hybrid network, and Transformer-based on the same platform. *Table I demonstrates that the recognition performance of ASHT-KD VPR is better than the sub-optimal model (R^2Former), and the average recall rate with $top-1$ and $top-5$ on the four datasets is increased by* ***1.275%*** *and* ***2.025%****, respectively.* In addition, The hybrid CNN and ViT VPR (hybrid CNN-Trans) method also achieves better performance than the traditional CNN methods.

D. Ablation Study

1) Ablation Studies with Different Teaching Modalities: In the training phase, the teacher model is trained under complex data samples such as low-illumination conditions, while the student models are trained under ideal data samples such as high-illumination conditions. In the place-matching phase, the Siamese student models with lightweight parameters are used for visual feature representation. We perform ablation studies based on different teaching modalities, i.e., single-student model, single-teacher model, ASHT-KD (hard-label teaching) model, ASHT-KD (soft-label teaching) model, and ASHT-KD (soft-hard label teaching) model.

We utilize multiple teachers (ASHT-KD with multiple teachers) to train two student models in a paradigm that combines soft and hard labels. *The results in Table II show that this method is better than using only a single teacher model with soft-hard label teaching (SH) to train two student models and is much more effective than the approach of using only soft*

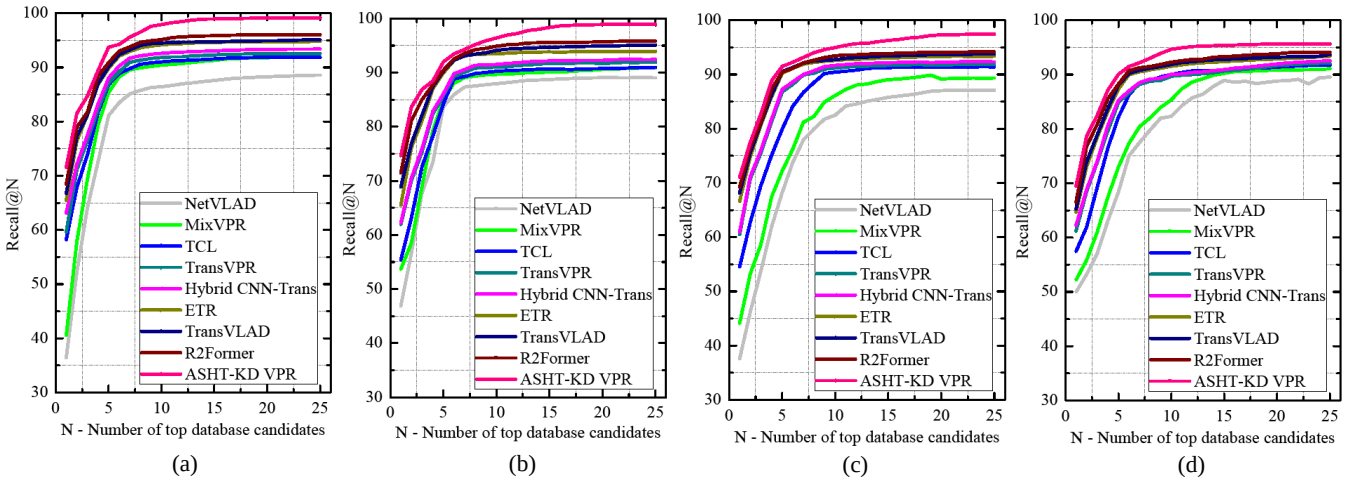


Fig. 4. Comparisons of the proposed ASHT-KD VPR with SOTA methods on four benchmarks: (a)-(d): KITTI_00, Tokyo 27/4, VPRICE, and Nordland.

TABLE I
OFFLINE EVALUATIONS COMPARED TO SOTA METHODS ON FOUR BENCHMARKS. ALL METHODS INVOLVED IN THE COMPARISON RANK $top-n$ ($n=1$ AND $n=5$) PLACES THROUGH THE GLOBAL FEATURE METHODS.

	Methods	KITTI_00		Nordland		VPRICE		Nordland	
		Top-1 (%)	Top-5 (%)	Top-1 (%)	Top-5 (%)	Top-1 (%)	Top-5 (%)	Top-1 (%)	Top-5 (%)
CNN	AP-GEM [37]	48.5	52.6	11.6	13.4	48.7	55.3	45.7	51.4
	DenseVLAD [30]	50.2	54.8	12.8	21.2	48.8	56.9	56.0	58.1
	NetVLAD [33]	55.4	62.8	13.4	21.7	50.2	60.7	58.9	61.5
	Patch-NetVLAD [37]	73.2	79.8	50.6	58.9	79.5	83.4	78.7	83.5
	DELG	75.4	81.4	58.7	64.5	81.6	85.5	79.4	84.6
	SPE-VLAD [39]	64.2	69.1	18.6	23.2	55.6	64.5	61.4	68.2
Hybrid	Hybrid CNN-Trans [16]	81.8	87.0	73.1	78.8	84.9	89.6	81.8	85.9
	TransHash [40]	72.5	77.9	48.8	59.1	77.5	81.7	76.2	81.1
	ETR [35]	78.5	82.4	51.4	60.0	75.3	80.4	78.3	83.4
	TransVPR [18]	81.5	86.8	72.6	78.7	84.9	88.7	82.6	86.8
ViT	TransVLAD [36]	83.4	87.5	73.4	80.4	84.3	89.0	82.5	90.0
	iTPN [41]	84.7	87.8	75.8	83.5	85.5	89.6	85.7	90.1
	R^2Former [17]	85.3	88.6	78.4	86.3	84.8	91.5	87.4	91.6
	ASHT-KD VPR (our)	**86.2**	**90.7**	**80.2**	**88.8**	**86.8**	**94.3**	**87.0**	**92.3**

labels, hard labels or no soft-hard label teaching (w/o SH). It also should be noted that the single-label training paradigm does not improve performance compared to a single teacher or student model, which indicates the superiority of the paradigm that combines soft and hard labels to guide student training.

2) Ablation studies with Different Scale Factors: The sampling method can be non-overlapping or overlapping, if it is non-overlapping, the sampling scale factor is 2, that is, for each additional layer, the resolution of the row is 1/2 of the original feature maps. *Fig. 5 shows the performance evaluation with the different scale factors. It can be seen that ASHT-KD VPR achieves the best results when the scale factor meets* $w = 2$.

3) Ablation studies with Different Number of Teacher Networks: From the results in Fig. 6 and Table III, we can see that increasing the number of teacher networks can improve recognition performance. However, it can also be observed that the performance is very limited after exceeding the baseline number (solid line in the figure). *Compared with the baseline (3 teacher networks), the recognition performance of the Top 1 candidate under the six teachers is improved by 2.9%, 1.9%, 1.9%, and 3.1% on the four datasets respectively. Therefore,*

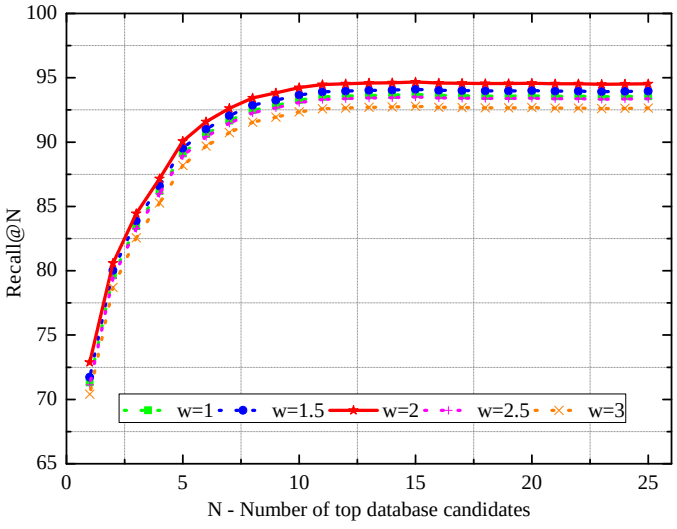


Fig. 5. Performance evaluation with different scale factors. To ensure the continuous and effective teaching and learning of the model in the process of knowledge distillation, we keep the scale factors of the feature coding in the teacher model and the student model consistent to ensure smoother knowledge transfer.

TABLE II
THE COMPARISON OF ALL TEACHING AND STUDENT MODELS. WHERE ASHT-KD (S), ASHT-KD (H), AND ASHT-KD (SH) REPRESENT THE SOFT LABEL, HARD LABEL, AND COMBINATION OF SOFT AND HARD LABEL TRAINING PARADIGMS. IT SHOULD BE NOTED THAT IN THIS EXPERIMENT WE USE 3 TEACHER NETWORKS TO TRAIN THE STUDENT MODEL.

Model	KITTI_00		Tokyo 24/7		VPRICE		Nordland	
	Top-1 (%)	Top-5 (%)	Top-1 (%)	Top-5 (%)	Top-1 (%)	Top-5 (%)	Top-1 (%)	Top-5 (%)
Single student (w/o SH)	85.3	88.2	78.4	86.8	85.4	93.0	84.8	89.1
Single teacher (w/o SH)	86.2	91.3	80.7	89.2	87.4	94.7	87.6	92.5
ASHT-KD (S) with multiple teacher	85.7	88.9	78.5	87.9	86.1	93.6	85.4	90.8
ASHT-KD (H) with multiple teacher	86.0	90.4	79.6	88.4	86.8	94.1	86.3	91.8
ASHT-KD (SH) with single teacher	86.2	90.7	80.2	88.8	86.8	94.3	87.0	92.3
ASHT-KD (SH) with multiple teacher	**86.9**	**91.4**	**81.0**	**89.5**	**88.1**	**95.2**	**87.8**	**92.9**

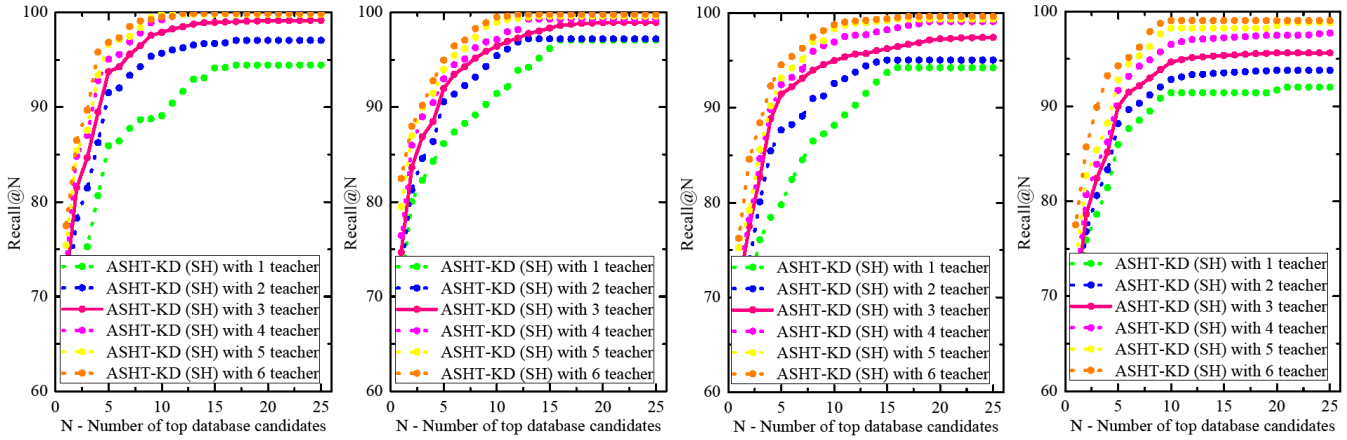


Fig. 6. Performance evaluation with different number of teacher networks on four benchmarks: (a)-(d): KITTI_00, Tokyo 27/4, VPRICE, Nordland.

TABLE III
RECALL@1 AND RECALL@5 WITH DIFFERENT NUMBERS OF TEACHER NETWORKS ON FOUR BENCHMARKS: (A)-(D): KITTI_00, TOKYO 27/4, VPRICE, NORDLAND.

Model	KITTI_00		Tokyo 24/7		VPRICE		Nordland	
	Top-1 (%)	Top-5 (%)	Top-1 (%)	Top-5 (%)	Top-1 (%)	Top-5 (%)	Top-1 (%)	Top-5 (%)
ASHT-KD (SH) with 1 teacher	80.8	85.5	73.8	82.1	81.6	89.4	81.9	86.7
ASHT-KD (SH) with 2 teacher	83.4	88.1	76.9	85.7	84.8	92.6	85.2	90.9
ASHT-KD (SH) with 3 teacher	86.2	90.7	80.2	88.8	86.8	94.3	87.0	92.3
ASHT-KD (SH) with 4 teacher	87.4	91.5	82.5	90.1	87.6	95.7	88.6	94.0
ASHT-KD (SH) with 5 teacher	88.7	93.1	84.0	92.5	88.5	96.2	89.4	94.8
ASHT-KD (SH) with 6 teacher	88.8	93.3	84.2	92.5	88.5	96.4	89.8	95.0
Δ	$\uparrow$ 2.9%	$\uparrow$ 2.8%	$\uparrow$ 4.8%	$\uparrow$ 4.0%	$\uparrow$ 1.9%	$\uparrow$ 2.2%	$\uparrow$ 3.1%	$\uparrow$ 2.8%

*so the average improvement is **3.18%**.*

4) Ablation studies with Different Skeleton Networks: To assess the effectiveness of the designed skeleton in feature-level distillation, we conduct ablation experiments using established networks as the skeleton in the distillation model. Specifically, VGG19 and VGG16 are employed as the teacher and student models (VGG19-VGG16), respectively. Other similar solutions include Res.50-Res.18, Res.101-Res.18. Moreover, the same DSFormer is designated as the teacher model and student (DSFor.-DSFor.), while the same APFormer is also used as the teacher and student model (APFor.-APFor.). Finally, the models APFormer and DSFormer proposed in this study are employed as the teacher and student models (ASHT-KD), respectively.

The results in Fig. 7 demonstrate that both the teacher and student models with the heavyweight APFormer to achieve superior performance. However, the computational complexity and inference speed are inferior to the ASHT solution. The other solutions exhibit slightly better computational complexity and inference speed than ASHT, but there is a significant performance gap compared to ASHT. Therefore, our ASHT is the most balanced solution within the current distillation framework.

E. Real-World Tests

The practicability of this framework is verified in a large number of real-world tests. The mobile robot takes a Jetson Xavier NX sever as the computing platform. The ASHT-KD place recognizer predicts the attention heatmaps in real-time

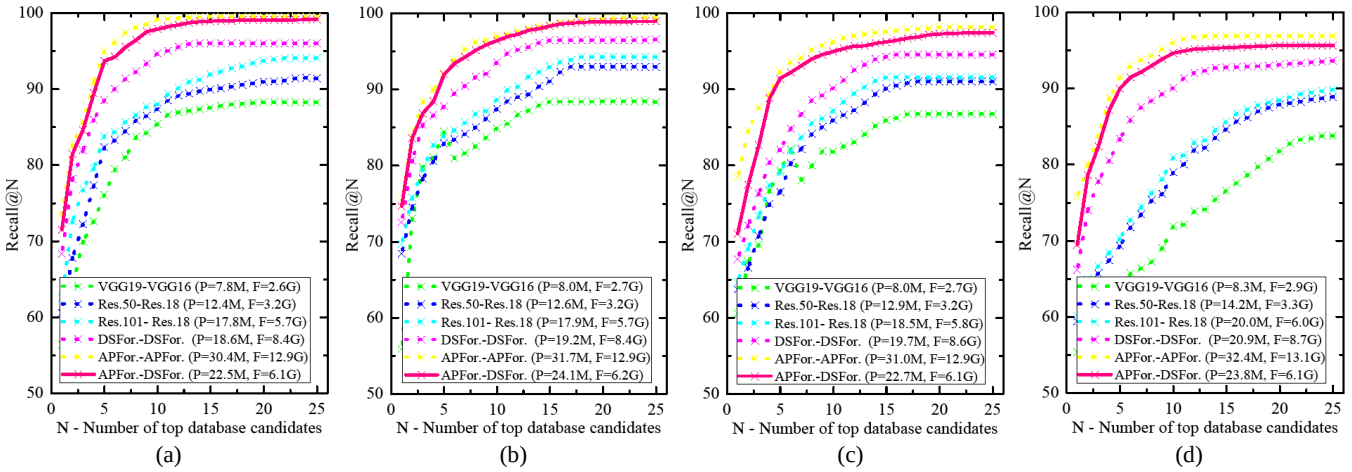


Fig. 7. Performance evaluation with different Skeleton networks including VGG, ResNet, ViTs, and ASHT-KD on four benchmarks: (a)-(d): KITTI_00, Tokyo 27/4, VPRICE, Nordland.

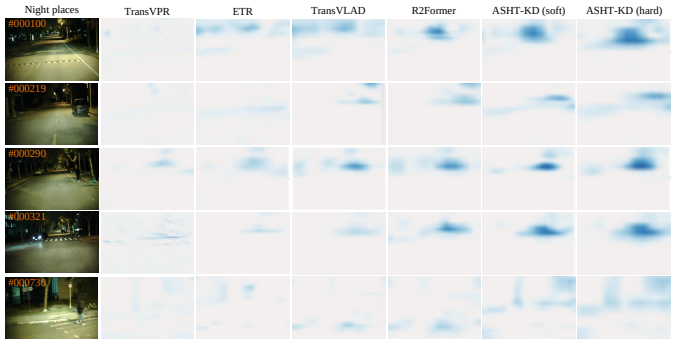


Fig. 8. Comparison of the attention feature heatmaps between ASHT-KD and SOTA ViT-based models at the testing. The ASHT-KD (soft) and ASHT-KD (soft) represent the soft-teaching student model and the hard-label teaching student model. The heatmaps are extracted through the SOTA VPR backbone networks.

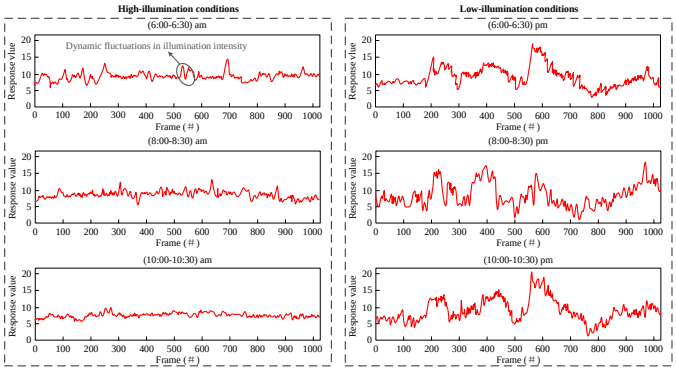


Fig. 9. The real-world VPR recognizer results and the fluctuation of attention activation values based on the RRLU function in high and low illumination conditions.

sending it to the mobile robot, helping the robot's automatic localization and navigation at night. Fig. 8 shows the recognition results of the attention heatmap and the comparison between the proposed ASHT-KD and SOTA VPR backbones. The ASHT-KD place recognizer performs better than other SOTA VPR backbones in poor illumination conditions. In addition, The comparison between the soft-label teaching student model and the hard-label teaching model shows the hard-label teaching KD-based paradigm can improve attention performance. *Overall, the hard-label teaching training paradigm in complex low-illumination can refine knowledge and efficiently transfer knowledge to the lightweight soft-label teaching training paradigm.*

Fig. 9 shows the real-world VPR recognizer results and the fluctuation of attention activation values based on the RRLU function. Our test experiment lasts from 6:00 am to 10:30 pm (six time periods). *The test results show that the average activation response values of the first to six stages are* 7.28, 7.12, 7.09, 7.75, 7.98, *and* 8.34, *respectively, which further confirms that our VPR framework has better recognition performance and strong generalization ability in complex environments, and can still maintain high-performance VPR under widely varying illumination conditions.*

Fig. 10 shows the average success rate of six tests (daytime scenes and nighttime scenes) in place recognition performance brought by the proposed three teaching-learning paradigms, i.e., soft-label teaching student-based VPR paradigm, hard-label teaching student-based VPR paradigm, and soft-hard label student-based VPR paradigm (ASHT-KD VPR). *The results demonstrate that knowledge distillation promotes the mobile robot VPR performance, especially in complex low-illumination environments, the improvement effect is more obvious.*

Table IV shows the real-time VPR performance (overall feature matching rate & overall place recognition rate) on real-world testing for different teaching-learning paradigms (soft-label teaching student paradigm, hard-label-teaching paradigm, and soft-hard label teaching paradigm). *In comparison with the soft-label teaching student paradigm, The ASHT-KD paradigm achieves* ***11.97%*** *and* ***6.8%*** *improvements in feature matching and position matching for daytime scenes, and* ***25%*** *and* ***31%*** *improvements in feature matching and place matching for*

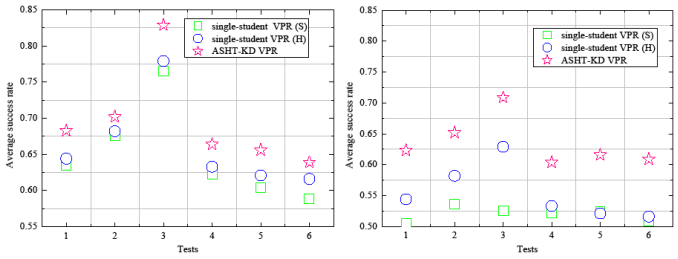


Fig. 10. The average success rate of six tests in place recognition performance brought by the proposed three teaching-learning paradigms. (a): daytime scenes (b): nighttime scenes.

TABLE IV
REAL-TIME VPR PERFORMANCE ON REAL-WORLD TESTING FOR DIFFERENT TEACHING-LEARNING PARADIGMS BASED ON DAYTIME SCENES AND NIGHTTIME SCENES.

	Teaching-Learning paradigms	Overall feature matching rate	Overall place recognition rate
Daytime scenes	Single-student (S)	0.794	0.6485
	Single-student (H)	0.854 (↑ 7.02%)	0.6625 (↑ 2.0%)
	ASHT-KD	0.902 (↑ 11.97%)	0.6955 (↑ 6.8%)
Nighttime scenes	Single-student (S)	0.664	0.519
	Single-student (H)	0.774 (↑ 14%)	0.578 (↑ 10%)
	ASHT-KD	0.889 (↑ 25%)	0.753 (↑ 31%)

nighttime scenes.

Examples of real-time test results in real-world environments are shown in Fig. 11. In challenging complex real-world experiments, compared with other SOTA place recognizers, our ASHT-KD recognizer maintains promising robustness and accuracy.

V. CONCLUSIONS

This paper proposes a novel adaptive soft-hard label teaching feature-level knowledge distillation learning framework (ASHT-KD) for all-day mobile robot VPR tasks. To be more specific, a multi-teacher teaching paradigm is constructed to guide the Siamese student network respectively based on the soft-label teaching paradigm and the hard-label teaching paradigm in directly extracting the low-illumination attention features for VPR matching. Different distillation methods are designed to teach different lightweight students, focusing on different emphases based on environmental properties. Teaching-learning feedback mechanism between students and teachers is built to encourage teachers and students to adjust their parameters at the same time to achieve an adaptive learning goal. ASHT-KD can reach the level of the teacher model on the whole. At the same time, ASHT-KD is faster than teacher mode because it is more lightweight. Furthermore, ASHT-KD achieves Superior performance compared with SOTA VPR_ers. Therefore, this work helps to maintain the speed of the model while maintaining good VPR performance and can implement soft-label teaching and hard teaching according to the complexity of the specific environment to achieve all-day visual place recognition of mobile robots.

Although the soft and hard label teaching knowledge distillation method has advantages in improving model performance, it also has drawbacks and challenges such as high computational resource consumption, strong dependence on teacher model quality, complex hyperparameter tuning, risk of information loss, and issues with model robustness. Future solutions include efficient model architecture design, automated hyperparameter tuning, robust knowledge distillation methods, multi-modal knowledge distillation, dynamic weight adjustment, and combining ensemble learning with knowledge distillation. These measures will help overcome the current method's shortcomings and further enhance its practicality and effectiveness.

REFERENCES

[1] S. Garg, N. Suenderhauf, and M. Milford, "Semantic–geometric visual place recognition: a new perspective for reconciling opposing views," The International Journal of Robotics Research, vol. 41, no. 6, pp. 573–598, 2022.
[2] K. A. Tsintotas, L. Bampis, and A. Gasteratos, "The revisiting problem in simultaneous localization and mapping: A survey on visual loop closure detection," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 11, pp. 19 929–19 953, 2022.
[3] X. Zhang, L. Wang, and Y. Su, "Visual place recognition: A survey from deep learning perspective," Pattern Recognition, vol. 113, p. 107760, 2021.
[4] Z. Zhu, K. Lin, A. K. Jain, and J. Zhou, "Transfer learning in deep reinforcement learning: A survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.
[5] J. Y.-L. Chan, K. T. Bea, S. M. H. Leow, S. W. Phoong, and W. K. Cheng, "State of the art: a review of sentiment analysis based on sequential transfer learning," Artificial Intelligence Review, vol. 56, no. 1, pp. 749–780, 2023.
[6] S. Niu, Y. Liu, J. Wang, and H. Song, "A decade survey of transfer learning (2010–2020)," IEEE Transactions on Artificial Intelligence, vol. 1, no. 2, pp. 151–166, 2020.
[7] L. Wang and K.-J. Yoon, "Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 6, pp. 3048–3068, 2021.
[8] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," International Journal of Computer Vision, vol. 129, pp. 1789–1819, 2021.
[9] F. Yuan, L. Shou, J. Pei, W. Lin, M. Gong, Y. Fu, and D. Jiang, "Reinforced multi-teacher selection for knowledge distillation," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 16, 2021, pp. 14 284–14 291.
[10] A. Chawla, H. Yin, P. Molchanov, and J. Alvarez, "Data-free knowledge distillation for object detection," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 3289–3298.
[11] C. Yang, H. Zhou, Z. An, X. Jiang, Y. Xu, and Q. Zhang, "Cross image relational knowledge distillation for semantic segmentation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12 319–12 328.
[12] X. Dong, J. Shen, F. Porikli, J. Luo, L. Shao. "Adaptive siamese tracking with a compact latent network", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, vol. 45, no. 7, pp. 8049-8062.
[13] W. Han, X. Dong, F. S. Khan, L. Shao, J. Shen. "Learning to Fuse Asymmetric Feature Maps in Siamese Trackers", In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16570-16580.
[14] K. Ying, Q. Zhong, W. Mao, Z. Wang, H. Chen, L. Y. Wu, Y. Liu, C. Fan, Y. Zhuge, C. Shen. "Ctvis: Consistent training for online video instance segmentation", In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 899-908.
[15] B. Li, L. Y. Wu, D. Liu, H. Chen, Y. Ye, X. Xie, "Image Template Matching via Dense and Consistent Contrastive Learning," 2023 IEEE International Conference on Multimedia and Expo (ICME), 2023, pp. 1319-1324.
[16] Y. Wang, Y. Qiu, P. Cheng, and J. Zhang, "Hybrid cnn-transformer features for visual place recognition," IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 3, pp. 1109–1122, 2022.
[17] S. Zhu, L. Yang, C. Chen, M. Shah, X. Shen, and H. Wang, "R2former: Unified retrieval and reranking transformer for place recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19 370–19 380.

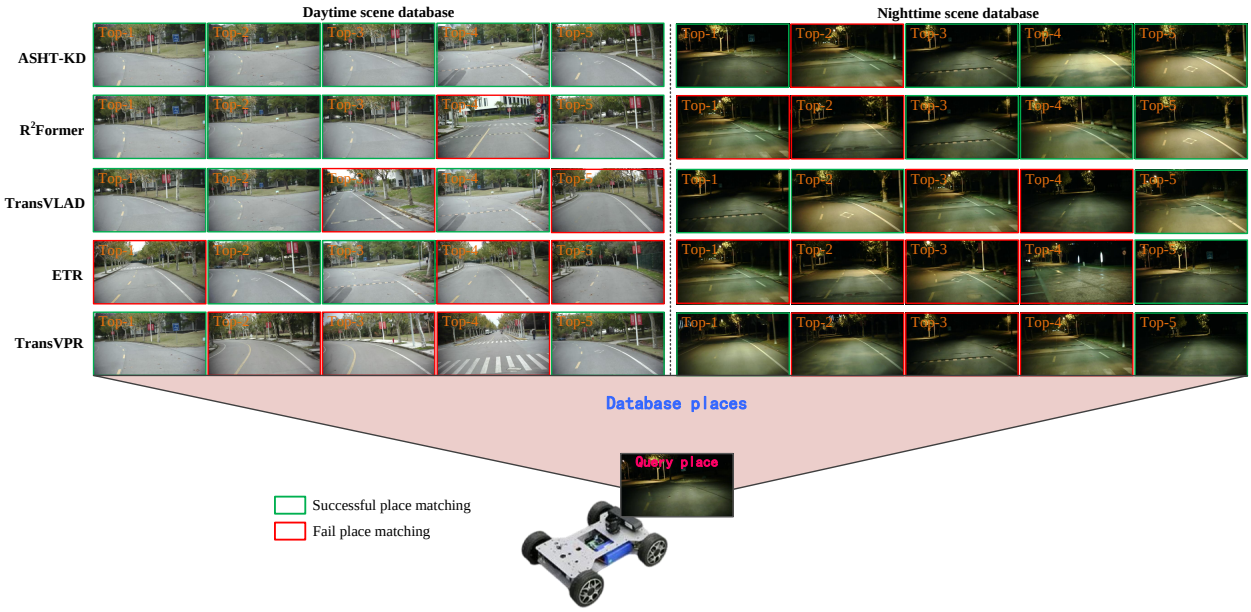


Fig. 11. Some place matching results of SOTA backbones on the real-world environments, where the green box represents the successful matching places and the red box represents the failed matching places. To verify the robustness in low-light conditions, we use the previously navigated daytime scenes and nighttime scenes as the currently retrieved robot position database, respectively.

[18] R. Wang, Y. Shen, W. Zuo, S. Zhou, and N. Zheng, "Transvpr: Transformer-based place recognition with multi-level attention aggregation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13 648–13 657.

[19] A. Ali-Bey, B. Chaib-Draa, and P. Giguere, "Mixvpr: Feature mixing for visual place recognition," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 2998–3007.

[20] D. Guo, H. Wang, and M. Wang, "Context-aware graph inference with knowledge distillation for visual dialog," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 10, pp. 6056–6073, 2021.

[21] Y. Feng, X. Sun, W. Diao, J. Li, and X. Gao, "Double similarity distillation for semantic image segmentation," IEEE Transactions on Image Processing, vol. 30, pp. 5363–5376, 2021.

[22] W. Chen, Y. Liu, N. Pu, W. Wang, L. Liu, and M. S. Lew, "Feature estimations based correlation distillation for incremental image retrieval," IEEE Transactions on Multimedia, vol. 24, pp. 1844–1856, 2021.

[23] S. Lin, H. Xie, B. Wang, K. Yu, X. Chang, X. Liang, and G. Wang, "Knowledge distillation via the target-aware transformer," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10 915–10 924.

[24] Z. Cai, X. He, J. Sun, and N. Vasconcelos, "Deep learning with low precision by half-wave gaussian quantization," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5918–5926.

[25] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4320–4328.

[26] T. Van Erven and P. Harremos, "R´enyi divergence and kullback-leibler divergence," IEEE Transactions on Information Theory, vol. 60, no. 7, pp. 3797–3820, 2014.

[27] E. Englesson and H. Azizpour, "Generalized jensen-shannon divergence loss for learning with noisy labels," Advances in Neural Information Processing Systems, vol. 34, pp. 30 284–30 297, 2021.

[28] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in International Conference on Learning Representations, 2018.

[29] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," The International Journal of Robotics Research, vol. 32, no. 11, pp. 1231–1237, 2013.

[30] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1808–1817.

[31] T. Naseer, M. Ruhnke, C. Stachniss, L. Spinello, and W. Burgard, "Robust visual slam across seasons," in 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2015, pp. 2529–2535.

[32] N. S¨underhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," in IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, 2015, pp. 4297–4304.

[33] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5297–5307.

[34] Y. Shen, R. Wang, W. Zuo, and N. Zheng, "Tcl: Tightly coupled learning strategy for weakly supervised hierarchical place recognition," IEEE Robotics and Automation Letters, vol. 7, no. 2, pp. 2684–2691, 2022.

[35] H. Zhang, X. Chen, Z. Y. Jing, Heming, Y. Wu, and C. Jin, "Etr: An efficient transformer for re-ranking in visual place recognition," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 5665–5674.

[36] H. Li, L. Zhang, D. Zhang, L. Fu, P. Yang, and J. Zhang, "Transvlad: Focusing on locally aggregated descriptors for few-shot learning," in European Conference on Computer Vision. Springer, 2022, pp. 524–540.

[37] J. Revaud, J. Almaz´an, R. S. Rezende, and C. R. d. Souza, "Learning with average precision: Training image retrieval with a listwise loss," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5107–5116.

[38] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "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, 2021, pp. 14 141–14 152.

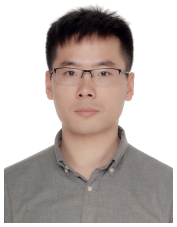
[39] J. Yu, C. Zhu, J. Zhang, Q. Huang, and D. Tao, "Spatial pyramid-enhanced netvlad with weighted triplet loss for place recognition," IEEEtransactions on Neural Networks and Learning Systems, vol. 31, no. 2, pp. 661–674, 2019.

[40] Y. Chen, S. Zhang, F. Liu, Z. Chang, M. Ye, and Z. Qi, "Transhash: Transformer-based hamming hashing for efficient image retrieval," in Proceedings of the 2022 International Conference on Multimedia Retrieval, 2022, pp. 127–136.

[41] Y. Tian, L. Xie, Z. Wang, L. Wei, X. Zhang, J. Jiao, Y. Wang, Q. Tian, and Q. Ye, "Integrally pre-trained transformer pyramid networks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 610–620.

Zhenyu Li (Member, IEEE) received the Ph.D. degree in Mechanical Engineering from Tongji University, Shanghai, China, in 2023. He is currently a lecturer with the Qilu University of Technology and also is currently an assistant researcher with the Shandong Academy of Sciences. His current research interests include intelligent perception, visual localization and navigation for robot automation in complex environments. He won the "Best Paper Finalist" in the 2019 IEEE-ROBIO Conference Selection. He has published more than 30 papers and served as the reviewer for several IEEE Transactions journals, such as IEEE TII, IEEE TMECH, IEEE TGRS, etc.

Pengjie Xu received the B.S., M.S., and Ph.D. degrees from Shandong University of Technology, Qingdao University, and Tongji University, China in 2015, 2018, and 2023, respectively. Currently, he works as a postdoctoral fellow with the School of Mechanical Engineering, Shanghai Jiao Tong University, China. His research interests include machine learning and robotics systems.



Zhenbiao Dong is currently a associate professor at the Shanghai Institute of Technology. He obtained a Ph.D degree from Shanghai Jiaotong University, Shanghai, China, in 2019. His research mainly focuses on data processing, compute vision, and advanced manufacturing technology.

Ruirui Zhang graduated from at School of Mechanical Engineering, Northwestern Polytechnical University, currently works at Qilu University of Technology. Ruirui Zhang does research in Mechanical Engineering and Materials Engineering. The current project include machine learning for advanced manufacturing technology.

Zhaojun Deng received the Ph.D. degree in Mechanical Engineering from Tongji University, Shanghai, China, in 2023. He is currently a Postdoctoral Fellow, Tongji University. His research interests include machine vision and photoelectric measuring technology.