Spatial-Temporal Expert Learning for Video-based Person Re-identification

Xiaofei Hui[1], Pengfei Wang[1], Evan Ling[2], Dezhao Huang[2], Keng Teck Ma[2], Minhoe Hur[2], Jun Liu[1†]

[1] Singapore University of Technology and Design, Singapore
[2] Hyundai Motor Group Innovation Center in Singapore (HMGICS), Singapore
h_xf@outlook.com, j.liu81@lancaster.ac.uk

Abstract. Video-based person re-identification (Re-ID) aims to retrieve the same identity in the query video clips from the gallery video clips. To solve this problem, exploiting fine-grained features is of great importance, especially when discriminating identities that are similar in appearance. In this paper, we propose to enhance the ability to explore fine-grained information with a novel input-aware extendable expert module. Instead of updating the network parameters with every sample in the dataset, we aim to train the experts within specific subsets that only contain similar samples and promote their ability to exploit fine-grained information within these similar samples. To achieve this goal, we incorporate two mechanisms in this module: input-aware expert selection mechanism and spatial-temporal selection mechanism. The first mechanism dynamically activates a set of experts on subsets of similar samples, pushing the experts to exploit subtle differences between these similar samples, while the second one further increases their sensitivity to the fine-grained differences in spatial and temporal aspects and allows the experts to dynamically utilize them for different input samples. In addition, to facilitate the expert module, we design an extendable scheme that allows the module to flexibly add new experts when necessary. As a result, our method achieves outstanding performance on two large-scale datasets.

Keywords: Video-base person Re-ID · Fine-grained feature learning · Expert module.

1 Introduction

Given a query image or video, person re-identification (Re-ID) aims to determine whether the same identity has appeared in other cameras or in the same camera at a different time. As a core capability for intelligent video surveillance and smart city applications, it has attracted sustained attention in recent years. Compared with image-based Re-ID, video-based Re-ID can exploit richer appearance and temporal cues, making it more robust to occlusion, viewpoint change, and missing body parts [1–5]. Despite the promising progress of existing

† Corresponding Author (j.liu81@lancaster.ac.uk).

methods [1, 2, 5–7], video-based Re-ID remains challenging in real-world surveillance scenarios.

Particularly, in real-world surveillance scenarios, different identities often exhibit highly similar global appearance, making coarse-level features insufficiently discriminative. As illustrated in Fig.1, such cases can easily confuse existing models. In contrast, fine-grained cues, such as hair, shoes, backpacks, and shoulder bags, can provide critical evidence for distinguishing identities[2, 7]. For example, in Fig. 1(a), the persons in white shirts can be distinguished by their *hair* and *bags*, while the persons in black shirts can be differentiated by their *backpacks*.

Many approaches have been proposed to enhance fine-grained feature learning for video-based Re-ID [2, 7, 3, 6, 4, 9]. For example, several studies incorporate attention mechanisms to highlight informative regions [2, 3, 9, 6], while others exploit part-level cues, such as attributes and human semantic parsing, to capture fine-grained details [10–12]. Though these methods have shown promising results, they still have limitations. In particular, network parameters are typically optimized using all training samples jointly. Yet, as shown in Fig.1, there can be various fine-grained cues to discriminate similar samples, each of which may only appear in a small subset of samples. As a result, optimization over all samples tends to bias the model toward generalizable coarse patterns, rather than learning the subtle cues that might only apply to a small group of samples [13].

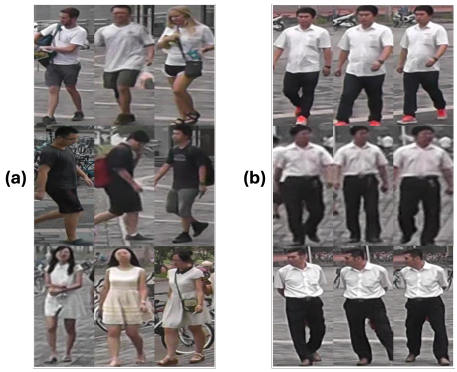

Fig. 1: Examples of different identities with similar looks sampled from MARS dataset [8]. (a) Single frames sampled from different identities that have subtle differences in appearance. Each row consists of three different identities. (b) Multiple frames sampled from three identities that look similar but have subtle differences in spatial and temporal aspects. Each row contains one identity.

Therefore, in order to improve the network's ability to learn the fine-grained cues, inspired by specialization learning mechanisms [14, 15], we develop an *input-aware extendable expert module* that adopts specialized parameters to particularly handle different subsets of similar samples. More concretely, we construct the expert module with a set of *experts*, each capable of learning information from the input samples. In this module, we apply an *input-aware expert selection mechanism* to activate the most relevant expert within a set of candidates, which is achieved by evaluating the relevance scores of each expert to the input with convolution operations. Naturally, input samples that are similar tend to have similar convolution results (i.e., expert relevance scores) and thus tend to activate the same experts [14–16]. In this way, the experts, activated with similar inputs, only need to focus on learning the fine-grained differences within

the particular subsets of similar samples, instead of being pushed to learn more general patterns that apply to more samples. Hence, each expert gains specialized power to exploit the subtle differences within a subset of similar samples, leading to improved performance in discriminating them. Moreover, we design an extendable scheme that allows the expert module to automatically add new experts during training. Intuitively, having too many experts in the expert module can be inefficient, and expert module of a small size may not be capable of handling large-scale datasets. With the extendable scheme, the expert module can automatically expand to fit its need, avoiding the trouble of finding a proper number of experts by handcrafted assigning or exhaustive search.

In addition, in video Re-ID, fine-grained information may lie in both spatial and temporal aspects. We may differentiate some samples using spatial fine-grained cues (e.g., shoes), while temporal information (e.g., gait and temporal pose information) can be useful to distinguish some other samples. As these subtle cues can exist more in either spatial or temporal aspect for different samples and may be easily neglected, it is crucial for the network to effectively learn and adaptively utilize them to discriminate different samples.

To achieve this goal, motivated by the spatial-temporal specialization in [14], we adopt the *input-aware spatial-temporal selection mechanism* in each expert to effectively and dynamically exploit the fine-grained differences in spatial and temporal aspects. After the expert is activated, we explicitly force each parameter to focus on either spatial or temporal aspect. As the expert is activated on a subset of similar samples with only fine-grained differences, this mechanism will further push each parameter to be more sensitive to subtle differences in a specific aspect. Meanwhile, the spatial-temporal selection forces the expert to dynamically focus on either spatial or temporal aspect for each input feature channel, allowing the experts to adaptively attach more importance to either spatial or temporal aspects for different samples. In this way, our expert module can effectively exploit fine-grained spatial-temporal information and achieve dynamic spatial-temporal feature learning.

In summary, our contributions are as follows: (1) To fully exploit fine-grained features for video-based Re-ID, we propose a flexible expert module. We force the experts to exclusively discriminate among subsets of similar input samples with an expert selection mechanism, and encourage the experts to learn fine-grained information. (2) To further effectively enhance spatial-temporal feature learning, we dynamically utilize spatial and temporal fine-grained features based on each input sample with a spatial-temporal selection scheme. (3) We show the effectiveness of our method by evaluating our method on MARS [8] and LS-VID [17]. With the help of the input-aware extendable expert module, we are able to reach outstanding performances on both datasets.

2 Related Work

Fine-grained feature learning. Fine-grained feature learning has been a long-standing problem. While humans can learn not only the significant information

but also the minor details of a sample, it is still a challenge for neural networks. The ability to learn fine-grained information is essential to tasks such as fine-grained image classification [18–20] and fine-grained action recognition [14, 21, 15], where the inter-class differences can be subtle. In person Re-ID, learning fine-grained features is also an important challenge, especially when differentiating different identities in similar appearance [22–24, 2, 3, 7] where commonly-used features in coarse level (e.g., color of clothes) may not be discriminative enough. Previous works in exploring fine-grained features can be roughly summarized into two categories: attention-based methods [2, 3, 7, 9] and part-level methods [9, 4, 10–12]. The attention-based methods train the network to focus on the most informative regions. For example, He et al. [2] propose Dense Attention that extracts hybrid information from both convolution neural network (CNN) and self-attention to learn the preference of fine-grained features. Hou et al. [3] propose to learn fine-grained features by applying several parallel attention modules that force the network to explore different regions in the input video sequence. On the other hand, part-level methods explore fine-grained features by leveraging local regions (e.g., sub-parts of the frames and body parts). For example, Zhu et al. [10] generate part-level features with a knowledge distillation paradigm that learns from both global and part-level patches. Zhu et al. [12] propose to learn human body parts and their belongings by human semantic parsing method.

While these methods are all insightful and achieve good performances, they may not be able to effectively explore the fine-grained information. As the parameters of networks are updated using every samples in the dataset, the loss will push them to learn general patterns to contribute to discriminating more identities, as opposed to the fine-grained cues that may only apply to subsets of samples [13]. Hence, from a new perspective, we intend to alleviate this problem by explicitly activating the experts on subsets of similar samples and pushing them to only focus on learning the subtle differences within the subsets of similar samples. In this way, our expert module gains enhanced ability to effectively learn the fine-grained differences between similar samples.

Spatial-temporal feature learning. In video-based Re-ID, leveraging spatial and temporal features is of great significance. There are different approaches [6, 3, 7, 4, 23, 25, 5, 26–28] to emphasize the importance of utilizing spatial and temporal information. For example, Aich et al. [7] propose spatial-temporal representation factorization to aggregate temporal and spatial features in high and low frequency. Eom et al. [26] introduce spatial and temporal memories to store spatial distractors and typical temporal patterns. Wang et al. [5] adopt a pyramid structure that progressively aggregates spatial and temporal features. Also, [4, 25] utilize graph convolutions to model the spatial-temporal information between frames. Different from previous methods, in this paper, we aim to improve the network's ability to learn fine-grained differences in spatial and temporal aspects by explicitly forcing each of the parameters to focus on one particular aspect. Because the expert only needs to focus on the fine-grained differences within a subset of similar samples, this mechanism will further push the parameters to

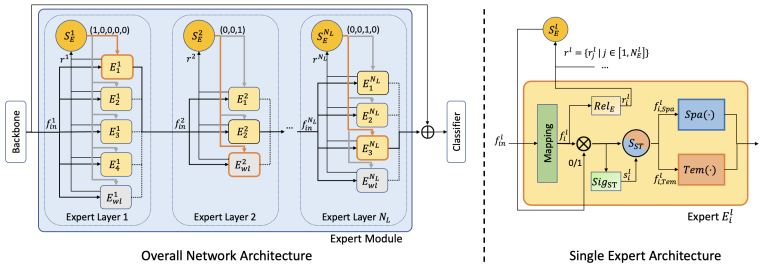

Fig. 2: Architecture of the proposed expert module. (Left) Each expert layer contains a number of experts E_i^l and an additional wait-list expert E_{wl}^l. For every input feature f_{in}^l, the expert selection mechanism S_E^l outputs a one-hot vector according to the relevance score vector r^l and activates one expert in each layer (indicated by orange arrows) while other experts are deactivated (indicated by gray arrows). Then the activated expert produces the output feature of the layer. A skip connection is then applied (indicated by $\bigoplus$). The classifier is only utilized during training. (Right) In the expert E_i^l, we first process the input feature f_{in}^l with a mapping procedure and obtain the mapped feature f_i^l. Then f_i^l is fed into the relevance evaluation module Rel_E to compute a relevance score r_i^l. The expert selector S_E^l in each layer takes relevance scores from every expert and produces a one-hot vector to activate or deactivate the expert (indicated by multiplying 0 or 1 with $\bigotimes$). Once the expert is activated, it further computes a spatial-temporal significance vector s_i^l with the spatial-temporal significance module Sig_{ST}. The spatial-temporal selector S_{ST} divides f_i^l into $f_{i,Spa}^l$ and $f_{i,Tem}^l$ according to s_i^l, and feed them into spatial and temporal branch respectively. We indicate spatial and temporal branch by $\mathrm{Spa}(\cdot)$ and $\mathrm{Tem}(\cdot)$.

be more sensitive to fine-grained cues in spatial and temporal aspects. Also, the dynamic selection allows the expert to adaptively exploit spatial and temporal information for different samples.

Dynamic network. As opposed to static networks, dynamic neural networks can adjust their structures or parameters according to the input, which adds to the representation power, adaptiveness, and interpretability [29]. Typical dynamic structures include dynamic depth [30], dynamic width [31], and dynamic routing [32]. Specifically, some previous works [14, 15] generate decisions for dynamic routing using expert modules with fixed structure to learn different human actions. Differently, we design an extendable expert module that can grow its capacity during training, enabling flexible and input-aware modeling of fine-grained spatial and temporal features. Our designs are particularly effective for distinguishing visually similar individuals in video-based person Re-ID.

3 Method

3.1 Overview

One major challenge in video-based person Re-ID is distinguishing different identities with highly similar appearance [2, 7]. In such cases, coarse appearance cues,

such as overall shape and clothing color, are often insufficient for reliable matching. However, most existing methods train network parameters using all samples in the dataset, which tends to encourage the learning of broadly shared patterns rather than subtle fine-grained differences that may only appear in small subsets of visually similar samples.

To address this issue, inspired by specialization learning in fine-grained modeling [14, 15], we propose an expert module that specifically learns fine-grained discrimination within subsets of *similar samples*. As shown in Fig. 2, the module contains N_L expert layers, each consisting of N_E^l experts and an expert selector ($l \in [1, N_L]$). Given the coarse-grained features extracted by a backbone network, the selector dynamically activates the most relevant expert in each layer according to the input. Since similar samples tend to yield similar relevance scores, they are likely to activate the same experts. As a result, each expert is updated on a subset of similar samples, encouraging it to capture subtle discriminative cues. To avoid manually specifying the number of experts, we further introduce an extendable scheme that automatically appends wait-list experts when needed. In addition, to better exploit fine-grained video cues, inspired by [14], we adopt parallel spatial and temporal branches in each activated expert. An input-aware spatial-temporal feature selection mechanism dynamically routes each feature channel to either the spatial or temporal branch, allowing the model to adaptively focus on the most informative subtle differences for each sample.

With these two dynamic selection mechanisms, the proposed module can effectively learn fine-grained spatial-temporal representations for video-based person Re-ID. Below we formally introduce the details of the proposed mechanisms.

3.2 Input-aware Extendable Expert Selection Mechanism

In each expert layer, we adopt a selection mechanism to dynamically match the input with the most relevant expert [14, 15]. Also, to flexibly construct the expert module to handle the samples, we design an extendable scheme that allows adding new experts during training.

More concretely, given a video clip, we first extract the coarse-grained features with the backbone network as the input to the expert module. We denote the input feature to the l-th expert layer as $f_{in}^l \in \mathbb{R}^{C\times T\times H\times W}$. In the i-th expert E_i^l, we process the input feature with the mapping module and obtain the mapped feature $f_i^l \in \mathbb{R}^{C\times T\times H\times W}$. Then the mapped feature f_i^l is fed into the relevance evaluation module Rel_E to evaluate the relevance of the expert to the input sample. Specifically, in Rel_E, we adopt max-pooling and a fully connected layer to obtain a relevance value, and normalize the value with tanh function to obtain the relevance score $r_i^l \in \mathbb{R}$:

$$\begin{aligned} f_i^l &= \mathrm{Mapping}(f_{in}^l), \\ f_{i,\max}^l &= \max_{T,H,W}(f_i^l),\ f_{i,\max}^l \in \mathbb{R}^{C\times 1}, \\ r_i^l &= \tanh(w_E^{i,l} f_{i,\max}^l), w_E^{i,l} \in \mathbb{R}^{1\times C}, \end{aligned} \tag{1}$$

where Mapping$(\cdot)$ represents the convolution operation in the mapping module, and $w_E^{i,l}$ denotes the linear transformation in the fully connected layer.

After obtaining relevance scores for every expert in the layer, the aggregated relevance vector $r^l = \{r_1^l, r_2^l, ..., r_{N_E^l}^l\}$ is fed into the expert selector S_E^l. Using the Gumbel-Softmax method [33, 14, 15], S_E^l evaluates r^l and generates a one-hot vector with the s-th value being 1. We then activate the s-th expert E_s^l and deactivate other experts in the layer. Notably, only the activated expert E_s^l is involved in the following process, and therefore during training, the parameters of each expert are only updated when activated.

In addition, we adopt an extendable expert scheme by keeping an additional expert in each layer (termed as wait-list expert E_{wl}^l) during training. For each input sample, the wait-list expert also generates a relevance score, and is only added to the layer if it has a higher relevance score to the input samples than the existing ones. Once the expert E_{wl}^l is appended to layer l, it becomes a constant expert $E_{N_L+1}^l$ and a new wait-list expert is generated automatically. The intuition behind this extendable scheme is that it empowers the network to adaptively adjust the number of experts and avoids the trouble of manually assigning the proper number of experts.

The expert selection mechanism plays an important role in improving the ability to learn fine-grained information. Most significantly, it achieves dynamic activation of the experts on subsets of similar samples. As the experts are only updated on subsets of similar samples instead of being trained with all samples, they are forced to focus on learning fine-grained differences within particular subsets, which promotes the ability to discriminate similar samples.

3.3 Input-aware Spatial-Temporal Selection Mechanism

After activating the expert for the input sample, we further explore spatial and temporal fine-grained features dynamically. Specifically, inspired by spatial-temporal fine-grained feature modeling [14], we construct two parallel branches that focus on spatial and temporal information respectively as shown in Fig.2. The spatial branch consists of a $1 \times 3 \times 3$ convolution layer to express spatial information, while the temporal branch contains a $3 \times 1 \times 1$ convolution layer that learns temporal information. Motivated by the observation that the discriminative information may lie more in either spatial or temporal aspects for different samples, we allow the experts to adaptively adjust their emphasis on the two aspects by forcing every feature channel to dynamically select between spatial and temporal branches for each input sample.

More concretely, we leverage the mapped feature f_s^l in the selected expert E_s^l and compute a significance vector $s_s^l \in \mathbb{R}^{C\times 1}$ with the spatial-temporal significance evaluation module Sig_{ST}, indicating whether the feature channel is more significant in spatial aspect or temporal aspect. In Sig_{ST}, we utilize max-pooling and fully connected layer to generate a vector $s_s^l \in \mathbb{R}^C$ that contains one

significance value for each feature channel:

$$\begin{aligned} f_{s,\max}^{l} &= \max_{T,H,W}(f_s^l), f_{s,\max}^{l} \in \mathbb{R}^{C\times 1}, \\ s_s^l &= \tanh(w_{ST}^{s,l} f_{s,\max}^{l}), w_{ST}^{s,l} \in \mathbb{R}^{C\times C}, \end{aligned} \tag{2}$$

where $w_{ST}^{s,l}$ represents the linear transformation in the fully connected layer.

After obtaining the significance vector s_s^l, we generate a binary decision vector $d_s^l \in \mathbb{R}^C$ with the help of the Improved Semhash method [34, 35]. Specifically, for the c-th feature channel, we obtain a decision value $d_{s,c}^l$. When $d_{s,c}^l = 1$, the feature slice $f_{s,c}^l \in \mathbb{R}^{1\times T\times H\times W}$ goes to the spatial branch; otherwise, it is fed into the temporal branch. The selection process is formulated as:

$$f_{s,Spa}^l = f_s^l \odot d_s^l, \quad f_{s,Tem}^l = f_s^l \odot (\mathbf{1} - d_s^l), \tag{3}$$

where $f_{s,Spa}^l$ and $f_{s,Tem}^l$ represent the input features to spatial and temporal branches respectively, $\mathbf{1}$ is a vector of 1's of size C, $\odot$ denotes multiplication in the channel dimension, and the elements in d_s^l and $(\mathbf{1} - d_s^l)$ are considered as channels for simplicity.

The output feature $f_{s,out}^l$ of the s-th expert in the l-th layer is obtained by adding the outputs from the two branches:

$$f_{s,out}^l = \mathrm{Spa}(f_{s,Spa}^l) + \mathrm{Tem}(f_{s,Tem}^l), \tag{4}$$

where $\mathrm{Spa}(\cdot), \mathrm{Tem}(\cdot)$ represent operations in the spatial and temporal branches respectively.

As the expert is forced to choose between spatial and temporal branches for each input feature channel, during training, it learns to adaptively assign the feature channels to the branch that can lead to greater discriminative capacity. In this way, the experts are able to effectively and efficiently explore fine-grained information in spatial and temporal aspects according to the input samples. By forcing the spatial and temporal branches to focus on their specialties, they are stimulated to be sensitive to subtle differences in their own specialties. Therefore, during inference, the experts are able to dynamically utilize spatial and temporal features and effectively exploit fine-grained information.

3.4 Loss Function

In our expert module, we encourage the experts to learn fine-grained features to identify different identities. Intuitively, the experts in the same layer need to be less similar to each other in order to handle different samples in the dataset. To achieve this goal, we additionally adopt a diversity loss $\mathcal{L}_{div}$ to limit the pair-wise similarity among the experts. More concretely, for each expert, we first obtain p_i^l and q_i^l by vectorizing the parameters in the spatial and temporal branches respectively, and compute the pair-wise cosine similarity of the vectorized spatial and temporal parameters within the same layer:

$$\mathcal{L}_{div}^{l,Spa} = \sum_{i=1}^{N_E} \sum_{j=1,j\neq i}^{N_E} \frac{{p_i^l}^\top p_j^l}{\|p_i^l\|_2 \|p_j^l\|_2}, \quad \mathcal{L}_{div}^{l,Tep} = \sum_{i=1}^{N_E} \sum_{j=1,j\neq i}^{N_E} \frac{{q_i^l}^\top q_j^l}{\|q_i^l\|_2 \|q_j^l\|_2}. \tag{5}$$

We further aggregate the diversity loss across all layers by summation:

$$\mathcal{L}_{div} = \frac{1}{2N_L}\sum_{l=1}^{N_L}\mathcal{L}_{div}^{l,Spa} + \frac{1}{2N_L}\sum_{l=1}^{N_L}\mathcal{L}_{div}^{l,Temp}. \tag{6}$$

In this way, a small $\mathcal{L}_{div}$ indicates that the parameters of the experts are nearly orthogonal to each other, so they share minimal common knowledge. During training, the diversity loss will penalize the pair-wise similarity between the experts, which helps the network to exploit diverse fine-grained information.

Overall, we train the network with the combination of cross entropy loss $\mathcal{L}_{ce}$, batch hard triplet loss $\mathcal{L}_{tri}$ [36], and $\mathcal{L}_{div}$:

$$\mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}_{tri} + \lambda\mathcal{L}_{div}, \tag{7}$$

where λ is a hyperparameter to balance the influence of diversity loss.

4 Experiments

Datasets. We carry out experiments to evaluate the performance of our proposed methods on two large-scale video-based person Re-ID datasets: MARS [8] and LS-VID [37]. Mars [8] is one of the largest video-based person Re-ID datasets, consisting of 17,503 sequences captured by six cameras. There are 1,261 identities in total, with 625 identities in the training set and 636 identities in the test set. LS-VID [37] dataset is another large-scale dataset for video-based person Re-ID captured by 3 indoor cameras and 12 outdoor cameras. This dataset consists of 14,943 sequences of 3,772 pedestrians. There are 842 identities in the training set, 200 identities in the validation set, and 2,730 identities in the test set.

Evaluation Metrics. Following previous person Re-ID methods [1–3, 5], we adopt Cumulated Matching Characteristics (CMC) curve and mean Average Precision (mAP) as evaluation metrics.

4.1 Implementation Details

Following SINet [1], we use ResNet-50 [38] as the backbone. The expert module is initialized with two experts ($N_E^l = 2$) for three layers ($N_L = 3$). In each expert, the mapping module consists of a $1 \times 1 \times 1$ convolution layer, followed by batch normalization and ReLU. The spatial branch contains a $1 \times 3 \times 3$ convolution, batch normalization and ReLU, while the temporal branch consists of a $3 \times 1 \times 1$ convolution, followed by batch normalization and ReLU. We apply $1 \times 1 \times 1$ convolution to reduce the feature channels before appending the expert module. The dimensions of the features (C, T, H, W) are dependent on the backbone network. We set the hyperparameter λ in Eq.7 to 0.1. During training, we adopt the Restricted Random Sampling (RRS) strategy [39] and sample 4 frames for each tracklet. Following [1, 5], the training batch is constructed with 8 different identities, each including 4 tracklets. The network is trained on a single RTX 3090 GPU. We use the Adam optimizer with learning rate of 0.0005 and weight decay of 0.0005. During inference, we split the videos into clips of 4 frames and adopt cosine similarity to measure the distance between the query and the gallery. We

Table 1: Comparison of our method with state-of-the-art video-based person ReID methods on MARS and LS-VID.

Methods	MARS				LS-VID	
	mAP	rank-1	rank-5	rank-20	mAP	rank-1
MGRA [9]	85.9	88.8	97.0	98.5	-	-
STGCN [4]	83.7	89.9	-	-	-	-
TCLNet-tri [23]	85.1	89.8	-	-	-	-
BiCnet-TKS [3]	86.0	90.2	-	-	75.1	84.6
GRL [6]	84.8	91.0	96.7	98.4	-	-
CTL [25]	86.7	91.4	96.8	98.5	-	-
STRF [7]	86.1	90.3	-	-	-	-
DenseIL [2]	87.0	90.8	97.1	98.8	-	-
STMN [26]	84.5	90.5	-	-	69.2	82.1
PSTA [5]	85.8	91.5	-	-	-	-
SINet [1]	86.2	91.0	-	-	79.6	87.4
CAViT [40]	**87.2**	90.8	-	-	79.2	**89.2**
DSANet [41]	86.6	91.1	-	-	75.5	85.1
MIRE-GRAR [42]	86.6	91.5	96.8	98.5	75.5	85.1
EGMDL [43]	86.6	91.1	-	-	-	-
Ours	87.0	**91.6**	**97.4**	**98.9**	**81.0**	88.3

add noises sampled from standard Gaussian distribution to Improved Semhash during training, and no noises are added during inference. The temperature in Gumbel-Softmax [33] is set to 1.

4.2 Performance Comparison

We compare the performance of our proposed method with state-of-the-art methods [44, 9, 4, 23, 3, 6, 25, 7, 2, 26, 5, 1] on two large-scale datasets: MARS [8] and LS-VID [17]. The results are shown in Tab.1. On both datasets, we initialize the three expert layers with two experts and one wait-list expert, and allow the expert module to automatically append new experts to each layer. For MARS, the expert module stabilizes with expert number of $N_E^1, N_E^2, N_E^3 = 4, 2, 4$, while on LS-VID the expert module expands to a size of $N_E^1, N_E^2, N_E^3 = 4, 4, 4$.

On MARS, our proposed method achieves 87.0% in mAP and 91.6% in rank-1, achieving state-of-the-art performance. Specifically, compared with other methods exploring fine-grained features [9, 4, 3, 2, 6, 7], our method reaches top performance. This shows that our expert module has superior ability to learn useful information that can help improve performance. On LS-VID, our expert module also shows outstanding performance on both mAP and rank-1 metrics. Although CAViT [40] also shows competitive performances, we argue that CAViT is based on transformer and samples 8 frames for each identity while our results are achieved based on CNN with 4 frames sampled for each identity due to limitation of GPU memory.

4.3 Ablation Studies

We conduct ablation studies to evaluate our designs on LS-VID.

Expert Selection Mechanism. To evaluate the importance of the expert selection mechanism S_E, we conduct two experiments without the expert selection mechanism as shown in Tab.2: 1) activating all

Table 2: Impact of the expert selection mechanism S_E.

Models	mAP	rank-1
ResNet-50	73.3	82.8
Expert Module w/o S_E (average weighting)	79.9	87.3
Expert Module w/o S_E (random activation)	79.6	87.2
Expert Module w/ S_E	81.0	88.3

experts and averaging the outputs (**average weighting**), and 2) randomly activating the experts (**random activation**). As shown, the performances drop when we disable the expert selection mechanism. Our model achieves better results than the experiment with **average weighting** suggesting that our method benefits from dynamically activating and pushing the experts to specifically focus on subsets of samples compared to simply adding more model parameters through the experts. Also, compared to **random activation** of the experts, we show that the expert module is more effective with the input-aware design that intends to only activate the experts on subsets of similar samples.

Spatial-temporal selection mechanism. We investigate the importance of the spatial-temporal selection mechanism S_{ST} as shown in Tab.3. We evaluate the following variants: 1) only using the spatial branch or the temporal branch in each expert (**spatial branch** and **temporal branch**), 2) replacing the two-branch design with a single branch containing a $3\times3\times3$ convolution layer with batch normalization and ReLU (**single branch**), 3) fusing the two branches by averaging their outputs (**average fusion**), and 4) replace the selection mechanism with random assignment (**random selection**). From the results in Tab.3 we can observe that only learning fine-grained information in one aspect is not sufficient. Also, our expert model outperforms all compared variants, demonstrating its effectiveness.

Table 3: Impact of the spatial-temporal selection mechanism S_{ST}.

Models	mAP	rank-1
Expert Module w/o S_{ST} (spatial branch)	79.5	87.3
Expert Module w/o S_{ST} (temporal branch)	79.4	87.1
Expert Module w/o S_{ST} (single branch)	80.0	87.4
Expert Module w/o S_{ST} (average fusion)	80.1	87.6
Expert Module w/o S_{ST} (random selection)	80.0	87.3
Expert Module w/ S_{ST}	81.0	88.3

Diversity loss. In our method, we employ the diversity loss $\mathcal{L}_{div}$ to encourage the experts in the same layer to be more diverse. To evaluate its impact, we compare the performance of the expert module with diversity loss using different λ. As shown in Tab.4, adding $\mathcal{L}_{div}$ in the training scheme improves rank-1 and mAP scores, and our method achieves the best performance when $\lambda = 0.1$. Thus we set $\lambda = 0.1$ in our experiments.

Table 4: Impact of the diversity loss.

Models	mAP	rank-1
$\lambda = 0$ (without $\mathcal{L}_{div}$)	80.1	87.3
$\lambda = 0.05$	80.6	87.9
$\lambda = 0.1$	81.0	88.3
$\lambda = 0.15$	80.3	87.7
$\lambda = 0.5$	80.0	87.6

Extendable scheme. To evaluate the design of the extendable scheme which automatically decides the number of experts in each layer, we manually fix the number of experts and compare the performances. In these experiments, we disable all wait-list experts and assign the expert module with $N_E = 2, 3, 4, 5$ experts per layer. We also keep other settings the same as the main experiment, i.e., we construct 3 layers of experts and enable the selection mechanisms and diversity loss. The results are shown in Tab.5. As we can see, the scores increase when N_E grows from 2 to 4, and the improvement tapers off when N_E exceeds 4. Notably, the model with expert number automatically assigned with the extendable scheme (i.e., $N_E = 4$) surpasses all other models, which shows the effectiveness of the extendable scheme.

Table 5: Number of experts.

N_E	mAP	rank-1
2	80.3	87.5
3	80.4	87.7
4	81.0	88.3
5	80.4	87.9

Number of layers. To evaluate the impact of the number of layers in the expert module, we further conduct experiments with different layers of experts.

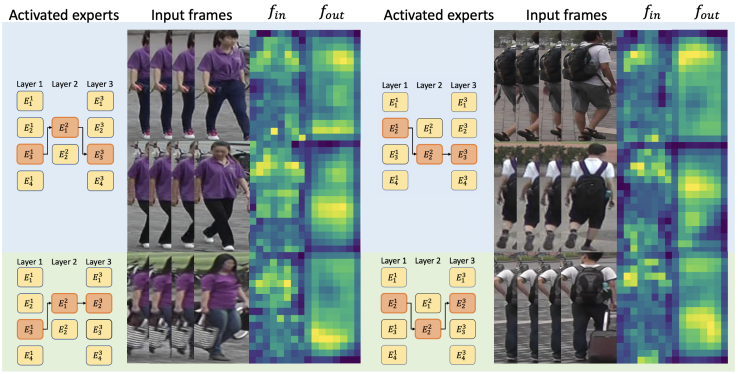

Fig. 3: Visualization of samples and features with the activated experts on MARS. The activated experts are indicated in orange, E_i^l represents the i-th expert in the l-th layer, f_{in} indicates the input feature to the expert module, f_{out} denotes the output feature of the expert module.

We construct the module with four experts per layer, and form the expert module with $N_L = 1, 2, 3, 5$ layer(s), as shown in Tab.6. As N_L grows from 1 to 3, the mAP increases by 1.3%. Yet, stacking too many layers (e.g., $N_L = 5$) doesn't bring improvement to the performance. Thus we set $N_L = 3$ in our experiments.

Table 6: Number of expert layers.

N_L	mAP	rank-1
1	79.7	87.3
2	80.4	87.5
3	81.0	87.9
5	80.9	87.8

Visualizations. To better evaluate the design of our expert module, we visualize some examples of the input samples and the activated experts in Fig.3. We show two groups of examples from MARS [8] with the input frames, input features going into the expert module, and the output features of the expert module. From the visualization results, we can observe that samples that look similar tend to have similar features and activate similar expert sets. While the input samples and input features are similar, our expert module can effectively learn discriminative information as shown by the different output features in Fig.3, demonstrating the effectiveness of our design.

5 Conclusion

In this paper, we propose an input-aware extendable expert module to tackle video-based Re-ID problem by exploiting the fine-grained discriminative details. We achieve this by dynamically activating the most relevant experts for the input samples, and adaptively adjusting the spatial-temporal feature learning according to the input sample. In this way, we encourage the network to exploit fine-grained spatial and temporal information. We demonstrate the effectiveness of our method with experiments on two large-scale video-based Re-ID datasets.

Privacy and ethical considerations. Video-based person re-identification relies on surveillance footage, raising privacy concerns. We use only publicly

available benchmarks for research purposes and do not identify real-world individuals. Notably, our expert module operates on abstract feature representations rather than raw images, and its modular design is naturally compatible with privacy-preserving techniques such as federated learning across cameras or adversarial feature perturbation to prevent identity leakage. We encourage responsible deployment with appropriate legal safeguards and oversight.

References

1. S. Bai, B. Ma, H. Chang, R. Huang, and X. Chen, "Salient-to-broad transition for video person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7339–7348.
2. T. He, X. Jin, X. Shen, J. Huang, Z. Chen, and X.-S. Hua, "Dense interaction learning for video-based person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 1490–1501.
3. R. Hou, H. Chang, B. Ma, R. Huang, and S. Shan, "Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 2014–2023.
4. J. Yang, W.-S. Zheng, Q. Yang, Y.-C. Chen, and Q. Tian, "Spatial-temporal graph convolutional network for video-based person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3289–3299.
5. Y. Wang, P. Zhang, S. Gao, X. Geng, H. Lu, and D. Wang, "Pyramid spatial-temporal aggregation for video-based person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 026–12 035.
6. X. Liu, P. Zhang, C. Yu, H. Lu, and X. Yang, "Watching you: Global-guided reciprocal learning for video-based person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 334–13 343.
7. A. Aich, M. Zheng, S. Karanam, T. Chen, A. K. Roy-Chowdhury, and Z. Wu, "Spatio-temporal representation factorization for video-based person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 152–162.
8. L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in *European conference on computer vision*. Springer, 2016, pp. 868–884.
9. Z. Zhang, C. Lan, W. Zeng, and Z. Chen, "Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 407–10 416.
10. K. Zhu, H. Guo, T. Yan, Y. Zhu, J. Wang, and M. Tang, "Pass: Part-aware self-supervised pre-training for person re-identification," in *European Conference on Computer Vision*. Springer, 2022, pp. 198–214.
11. K. Zhu, H. Guo, S. Zhang, Y. Wang, G. Huang, H. Qiao, J. Liu, J. Wang, and M. Tang, "Aaformer: Auto-aligned transformer for person re-identification," *arXiv preprint arXiv:2104.00921*, 2021.

12. K. Zhu, H. Guo, Z. Liu, M. Tang, and J. Wang, "Identity-guided human semantic parsing for person re-identification," in *European Conference on Computer Vision*. Springer, 2020, pp. 346–363.
13. J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, pp. 1–54, 2019.
14. T. Li, L. G. Foo, Q. Ke, H. Rahmani, A. Wang, J. Wang, and J. Liu, "Dynamic spatio-temporal specialization learning for fine-grained action recognition," in *European Conference on Computer Vision*. Springer, 2022, pp. 386–403.
15. L. G. Foo, T. Li, H. Rahmani, Q. Ke, and J. Liu, "Era: Expert retrieval and assembly for early action prediction," in *European Conference on Computer Vision*. Springer, 2022, pp. 670–688.
16. M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.
17. J. Li, J. Wang, Q. Tian, W. Gao, and S. Zhang, "Global-local temporal representations for video person re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3958–3967.
18. L. Yang, X. Li, R. Song, B. Zhao, J. Tao, S. Zhou, J. Liang, and J. Yang, "Dynamic mlp for fine-grained image classification by leveraging geographical and temporal information," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 10 945–10 954.
19. X. Yang, Y. Wang, K. Chen, Y. Xu, and Y. Tian, "Fine-grained object classification via self-supervised pose alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 7399–7408.
20. H. Touvron, A. Sablayrolles, M. Douze, M. Cord, and H. Jégou, "Grafit: Learning fine-grained image representations with coarse labels," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 874–884.
21. J. Munro and D. Damen, "Multi-modal domain adaptation for fine-grained action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
22. S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 013–15 022.
23. R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen, "Temporal complementary learning for video person re-identification," in *European conference on computer vision*. Springer, 2020, pp. 388–405.
24. P. Hong, T. Wu, A. Wu, X. Han, and W.-S. Zheng, "Fine-grained shape-appearance mutual learning for cloth-changing person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10 513–10 522.
25. J. Liu, Z.-J. Zha, W. Wu, K. Zheng, and Q. Sun, "Spatial-temporal correlation and topology learning for person re-identification in videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4370–4379.
26. C. Eom, G. Lee, J. Lee, and B. Ham, "Video-based person re-identification with spatial and temporal memory networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 036–12 045.
27. Y. Yan, J. Qin, J. Chen, L. Liu, F. Zhu, Y. Tai, and L. Shao, "Learning multi-granular hypergraphs for video-based person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2899–2908.

28. W. Wu, J. Liu, K. Zheng, Q. Sun, and Z.-J. Zha, "Temporal complementarity-guided reinforcement learning for image-to-video person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7319–7328.
29. Y. Han, G. Huang, S. Song, L. Yang, H. Wang, and Y. Wang, "Dynamic neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
30. X. Wang, F. Yu, Z.-Y. Dou, T. Darrell, and J. E. Gonzalez, "Skipnet: Learning dynamic routing in convolutional networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 409–424.
31. S. Cai, Y. Shu, and W. Wang, "Dynamic routing networks," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3588–3597.
32. G. E. Hinton, S. Sabour, and N. Frosst, "Matrix capsules with em routing," in *International conference on learning representations*, 2018.
33. E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=rkE3y85ee
34. L. Kaiser, S. Bengio, A. Roy, A. Vaswani, N. Parmar, J. Uszkoreit, and N. Shazeer, "Fast decoding in sequence models using discrete latent variables," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2390–2399.
35. Ł. Kaiser and S. Bengio, "Discrete autoencoders for sequence models," *arXiv preprint arXiv:1801.09797*, 2018.
36. A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.
37. J. Li, J. Wang, Q. Tian, W. Gao, and S. Zhang, "Global-local temporal representations for video person re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3958–3967.
38. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
39. C.-T. Liu, C.-W. Wu, Y.-C. F. Wang, and S.-Y. Chien, "Spatially and temporally efficient non-local attention network for video-based person re-identification," in *British Machine Vision Conference*, 2019.
40. J. Wu, L. He, W. Liu, Y. Yang, Z. Lei, T. Mei, and S. Z. Li, "Cavit: Contextual alignment vision transformer for video object re-identification," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV*. Springer, 2022, pp. 549–566.
41. M. Kim, M. Cho, and S. Lee, "Feature disentanglement learning with switching and aggregation for video-based person re-identification," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 1603–1612.
42. Z. Zhu, S. Chen, G. Qi, H. Li, and X. Gao, "Multi-granular inter-frame relation exploration and global residual embedding for video-based person re-identification," *Signal Processing: Image Communication*, vol. 132, p. 117240, 2025.
43. C. Cao, X. Fu, S. Xu, C. Ge, K. Wang, and Z.-J. Zha, "Learning robust event-guided representations for person re-identification: Cao et al." *International Journal of Computer Vision*, vol. 134, no. 2, p. 82, 2026.
44. X. Gu, B. Ma, H. Chang, S. Shan, and X. Chen, "Temporal knowledge propagation for image-to-video person re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9647–9656.