Vision Shared and Representation Isolated Network for Person Search

Yang Liu[1*], **Yingping Li**[2], **Chengyu Kong**[2], **Yuqiu Kong**[1], **Shenglan Liu**[1*] and **Feilong Wang**[1]
[1]School of Innovation and Entrepreneurship, Dalian University of Technology
[2]Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology
ly@dlut.edu.cn, {liyp, kongchengyu}@mail.dlut.edu.cn, {yqkong, liusl, wangfeilong}@dlut.edu.cn

Abstract

Person search is a widely-concerned computer vision task that aims to jointly solve the problems of pedestrian detection and person re-identification in panoramic scenes. However, the pedestrian detection focuses on the consistency of pedestrians, while the person re-identification attempts to extract the discriminative features of pedestrians. The inevitable conflict greatly restricts the researches on the one-stage person search methods. To address this issue, we propose a Vision Shared and Representation Isolated (VSRI) network to decouple the two conflicted subtasks simultaneously, through which two independent representations are constructed for the two subtasks. To enhance the discrimination of the re-ID representation, a Multi-Level Feature Fusion (MLFF) module is proposed. The MLFF adopts the Spatial Pyramid Feature Fusion (SPFF) module to obtain diverse features from the stem network. Moreover, the multi-head self-attention mechanism is employed to construct a Multi-head Attention Driven Extraction (MADE) module and the cascaded convolution unit is adopted to devise a Feature Decomposition and Cascaded Integration (FDCI) module, which facilitates the MLFF to obtain more discriminative representations of the pedestrians. The proposed method outperforms the state-of-the-art methods on the mainstream datasets.

1 Introduction

Person search is a computer vision task widely needed in the real world. It aims to search and locate target persons from the panoramic scenes in the gallery. Since [Xu *et al.*, 2014], person search has been continuously studied. Person search is related to pedestrian detection and person re-identification (re-ID) research, the former aims to detect all pedestrians in scenes, while the latter needs to retrieve target persons in hand-cropped pedestrian patches. Person search can also be seen as a combination of pedestrian detection and person re-ID, which is more suitable for real-world problems in open scenarios, such as suspect tracing and finding missing persons.

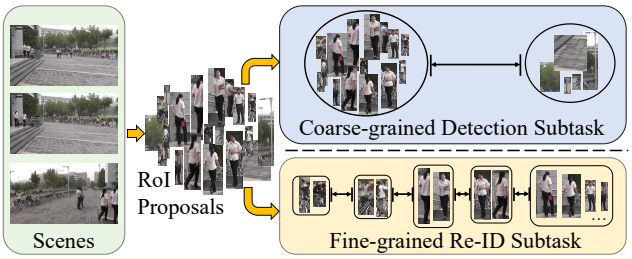
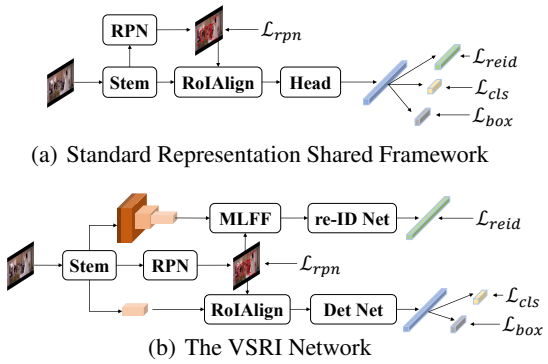


Figure 1: Illustration of the key challenges in the person search task. The pedestrian detection subtask treats all pedestrians as the same class, while the person re-identification subtask regards pedestrians with different identities as distinct classes.

To resolve the above two subtasks, an intuitive way is to employ a pedestrian detector to obtain pedestrian patches and then solve it as a person re-ID problem, which is a two-stage method. The other way is to jointly learn pedestrian detection and person re-ID in a unified framework, namely the one-stage method. Although two-stage methods can better benefit from the research in pedestrian detection and person re-ID, they require more parameters and computations. Hence, we pay more attention to one-stage methods.

For the one-stage method, a general way given by [Xiao *et al.*, 2017] is to attach an additional re-ID head based on Online Instance Matching (OIM) to Faster R-CNN [Ren *et al.*, 2015] detector. However, such a pipeline contains an inevitable conflict. As illustrated in Figure 1, pedestrian detection treats all persons as the same category, so all pedestrians are required to be as similar as possible in the feature space. Nevertheless, pedestrians with different identities are regarded as distinct classes in person re-ID, so feature representations with different identities are required to be dispersed as much as possible in the meanwhile. Standard one-stage methods allow detection subtask and re-ID subtask share the same feature representation, which leads to difficulties in joint optimization. To address this issue, recent works, such as [Dong *et al.*, 2020a] and [Zhang *et al.*, 2021], introduce knowledge distillation into OIM to strengthen the discrimination of representation. [Chen *et al.*, 2020b] decouples this conflict at the feature-level by respectively decomposing

*Corresponding Authors.



(a) Standard Representation Shared Framework

(b) The VSRI Network

Figure 2: Comparison between (a) standard representation shared framework and (b) our proposed VSRI network.

embedding into magnitude and angle for detection and re-ID. However, these methods do not change the core of the paradox that the model needs to acquire a shared representation to perform two conflicted subtasks. In contrast, we implement a simple but effective approach to decouple the conflicted subtasks at the network-level and obtain isolated task-specific feature representations from the shared region.

In this paper, we propose a Vision Shared and Representation Isolated (VSRI) network to tackle the aforementioned conflict. As shown in Figure 2, the vanilla one-stage method generally makes detection and re-ID to share the same feature representation, which results in the inherent optimization conflict. In contrast, our proposed method performs detection and re-ID through two independent branches. Considering that in CNNs, features in deep layers represent global semantic information and features in shallower layers tend to be more general, the proposed VSRI network allows two different branches to share the same feature map extracted by the stem network and the same RoI proposals proposed by the region proposal network (RPN). Still, each branch extracts its own feature representation containing further semantic information by an independent convolutional neural network. In general, our proposed VSRI network eliminates the inherent optimization conflict through a simple and intuitive way and retains all advantages of standard one-stage methods.

Compared with the detection subtask, the re-ID subtask requires more fine-grained information to distinguish pedestrians with different identities. For this reason, we design a Multi-Level Feature Fusion (MLFF) module to fuse shallow features into the final outputs of the stem network to obtain rich detailed information. Our designed MLFF uses the output feature maps of each ResNet [He *et al.*, 2016] stage as input. According to candidate regions given by RPN, these pyramidal feature maps are passed through the Spatial Pyramid Feature Fusion (SPFF) module to obtain the same size features. After channel-wise concatenation, we design a Multi-head Attention Driven Extraction (MADE) module, in which a Multi-Head Self-Attention (MHSA) mechanism [Vaswani *et al.*, 2017] is employed to estimate the global relationships between each position. Besides, we devise a Feature Decomposition and Cascaded Integration (FDCI) module to extract local features with different receptive fields by using Multi-Scale Convolution (MS-Conv) unit. Through this structure, features from different levels of the stem network are fully fused so that the subsequent re-ID branch can extract more discriminative feature representation from the fused feature to achieve better performance.

In conclusion, the contributions of this paper are three-fold. **First**, we propose a VSRI network to extract isolated task-specific feature representations from the shared region, which eliminates the natural conflict between pedestrian detection and person re-ID. To our knowledge, it is the first time to devise a vision shared and representation isolated architecture for the person search task. **Second**, we devise a complementarily integrated MLFF module to obtain the fine-grained representation, which consists of an information enhancement module (SPFF), a global feature extraction module (MADE), and a local feature integration module (FDCI). **Third**, experiments on real-world datasets demonstrate the superior performance of VSRI network in person search .

2 Related Work

Since [Xu *et al.*, 2014] raised the issue of person search, its rich application scenarios and practical significance have attracted numerous studies. The release of CUHK-SYSU [Xiao *et al.*, 2016] and PRW [Zheng *et al.*, 2017] has provided significant help for the research, promoting the development of this field. However, as early as [Xu *et al.*, 2014], it has been pointed out that there is a natural conflict between pedestrian detection and person re-ID.

Two-stage Methods. Two-stage methods implement pedestrian detection and person re-ID through two independent models successively to avoid the conflicts mentioned above. [Zheng *et al.*, 2017] first combines mainstream object detection methods and person re-ID methods to provide a performance benchmark for person search. [Lan *et al.*, 2018] improves the ability to detect different scales persons through cross-level semantic alignment loss. [Chen *et al.*, 2018] introduces an additional segmentation task to generate character masks so that the following re-ID model can focus on purer features. [Han *et al.*, 2019] proposed an RoI Transform Layer to connect both two models to build a joint training two-stage method. [Dong *et al.*, 2020b] draws on search-based methods then designs a query-guided framework to take advantage of query information. [Wang *et al.*, 2020] points out the consistency problem in two-stage methods and introduces an auxiliary task in the detector to alleviate this issue.

One-stage Methods. One-stage methods handle the two subtasks in a joint optimization approach. [Xiao *et al.*, 2017] first proposes a multi-task learning framework based on Faster R-CNN. With the help of additional lookup table and circular queue, OIM uses the mechanism of Online Instance Matching to optimize detection and re-ID simultaneously. In terms of deep representation learning, [Liu *et al.*, 2017] and [Chang *et al.*, 2018] respectively introduce long-short term memory and reinforcement learning into person search. [Yan *et al.*, 2019] improves search performance by constructing a graph with contextual clues. [Dong *et al.*, 2020a] designs BINet to reduce the interference of irrelevant information in the

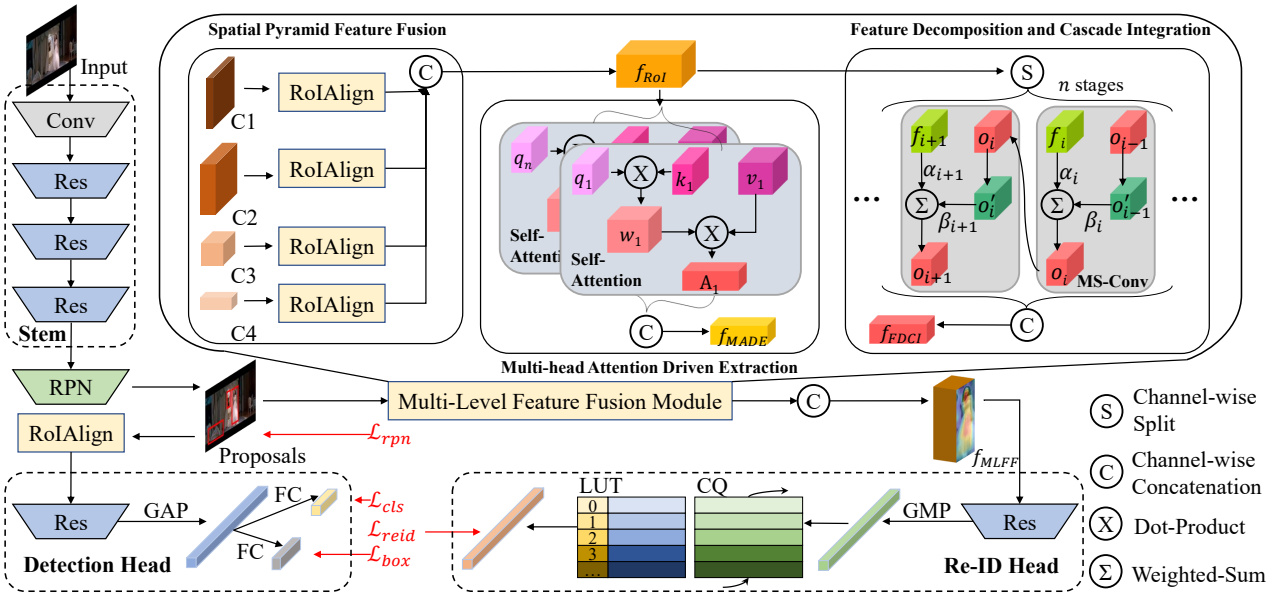


Figure 3: The overall architecture of our proposed VSRI network. VSRI network employs dual individual branches to obtain isolated feature representations for the conflicted subtasks in parallel. To enhance the discrimination of the re-ID representation, the MLFF, which consists of the SPFF, the MADE, and the FDCI module, is adopted.

receptive field. [Chen *et al.*, 2020b] decouples the two subtasks by decomposing features into angles and norms, representing identity information and classification confidence. As for deep metric learning, [Xiao *et al.*, 2019] introduces center loss into person search to shrink intra-class distance. [Chen *et al.*, 2020a] establishes a hierarchical relationship between detection and re-ID to obtain better embeddings. [Han *et al.*, 2021] proposed a decoupled pipeline based on a one-stage detector and supervised the learning process by pairwise loss.

In summary, the two-stage method is less efficient than the one-stage method and does not establish an association between the two subtasks. On the other hand, the one-stage method has to face the conflict between two subtasks caused by the shared representation. Therefore, we propose a VSRI network to improve the one-stage method by extracting isolated feature representations for the above conflicts while retaining the advantages of the one-stage method.

3 Method

The overall architecture of our proposed method is illustrated in Figure 3. In our proposed method, we adopt the VSRI network to improve the previous representation shared framework. At the same time, in the re-ID subtask, we design the MLFF to acquire and fuse multi-level features extracted from the stem network. Details of VSRI network and MLFF are described in the following sections.

3.1 Vision Shared and Representation Isolated Network

For one-stage methods, standard representation shared methods allow detection and re-ID to share the same features, which leads to inevitable conflicts. Therefore, we design the VSRI network as the basic framework in this paper.

First, we employ ResNet-50 pre-trained on ImageNet as a backbone model. The backbone is divided into two parts: the beginning to the end of conv4_x is called stem network, two identical conv5_x as follows, namely detection network and identification network. Given the input scene, the stem network first extracts the CNN feature. RPN predicts RoI proposals where pedestrians may exist based on this feature. After that, for each RoI proposal, the corresponding patch is sampled from the feature obtained from the stem network, then passes through the detection branch and the re-ID branch in parallel. In the detection branch, RoIAlign is adopted to sample the RoI feature from the whole scene feature. The sampled feature is further extracted via the detection network. GAP is employed to obtain the final feature representation. Two fully connected layers are following to perform fore-background classification and BBox regression. In the re-ID branch, the MLFF is utilized to sample and fuse multi-level features, while the identification network extracts high-level semantic features as follows. Moreover, we employ GMP to obtain re-ID features rather than GAP in the detection branch.

Overall, the VSRI network ensures two conflicting subtasks independently extract task-specific feature representations on a shared regional visual feature. Thus, the conflict in feature space in the traditional feature shared methods is avoided, and the information of two different subtasks can be used to assist the stem network to obtain better visual representation.

3.2 Multi-Level Feature Fusion

Since person re-ID needs to extract discriminative features to distinguish which of the many analogous identities the RoI belongs to, and there are also a large number of unlabeled pedestrians interference, re-ID subtask needs more fine-

grained feature representations to differentiate persons with different identities better. Considering that the output feature map of the stem network is still inevitably affected by the detection branch, only using this feature map can not describe a person's identity-specific information well. Therefore, we design the MLFF to take advantage of the output feature maps on each stage of the stem network, hoping to integrate shallow features containing detailed information into the output of the stem network as much as possible. In this way, the identification network can extract more discriminative fine-grained features.

Our designed MLFF includes the following three components. We utilize SPFF to obtain the feature f_{RoI} for each RoI proposal. The MADE is employed to mine global relationships from f_{RoI}, while the FDCI is adopted to extract local information of different scales from f_{RoI}. The final output f_{MLFF} is obtained by applying channel-wise concatenation on f_{MADE} and f_{FDCI}, i.e., f_{MLFF} can be described as

$$f_{MLFF} = \text{Concat}(MADE(f_{RoI}), FDCI(f_{RoI})) \quad (1)$$

Specifically, details of the three parts of the MLFF are introduced as follows.

Spatial Pyramid Feature Fusion (SPFF)

Considering that more detailed information is retained in the shallow features, which is crucial for us to extract the distinctive features of pedestrians. Therefore, we design the SPFF to obtain and fuse RoI features at multiple levels of the stem network. The stem network contains four parts, namely conv1, conv2_x, conv3_x, and conv4_x. We denote the outputs on the four stages of the stem network as $F_{C1}, F_{C2}, F_{C3}, F_{C4}$. To concatenate these features, we apply RoIAlign on feature maps of each level and output features with size 14×7, denoted as $f_{C1}, f_{C2}, f_{C3}, f_{C4}$. After that, a channel-wise concatenation layer is utilized to compose feature maps from different levels. The fused features are recorded as f_{RoI}, that is

$$f_{RoI} = \text{Concat}(f_{C1}, f_{C2}, f_{C3}, f_{C4}) \quad (2)$$

Multi-head Attention Driven Extraction (MADE)

In order to distinguish the human body from the background and further weaken the interference of irrelevant background information on person re-ID, we design a MADE module based on MHSA. To explore the global context, MADE employs the self-attention operation to reorganize values by computing the similarity of different keys and values generated from f_{RoI}.

Specifically, for each $f_{RoI} \in \mathbb{R}^{c\times h\times w}$ after performing layer normalization, we employ three 1×1 convolutional layers to project it to different feature spaces and reshape the results to $q_{RoI} \in \mathbb{R}^{(h\cdot w)\times \tilde{c}}, k_{RoI} \in \mathbb{R}^{(h\cdot w)\times \tilde{c}}, v_{RoI} \in \mathbb{R}^{(h\cdot w)\times c}$ as queries, keys, and values. After that, calculate the attention map according to Eq. 3 and resize it to the size of $c \times h \times w$.

$$\text{A} = \text{Softmax}(\frac{q_{RoI} \cdot k_{RoI}^{\top}}{\sqrt{\tilde{c}}}) \cdot v_{RoI} \quad (3)$$

This operation performs weighted summation on values according to the similarity among the queries and the keys corresponding to each element. After that, each value of the output attention map contains the global context. To fully mine global relationships among each element, we adopt MHSA to improve the effectiveness of the MADE. In MHSA, each head corresponds to a projection and performs a complete scaled dot-product self-attention process. After that, a channel-wise concatenation layer is utilized to compose the attention maps of all heads. That is, MADE can be described by Eq. 4, where n represents the number of heads.

$$MADE(f_{RoI}) = \text{Concat}(\text{A}_1, \text{A}_2, \cdots, \text{A}_n) \quad (4)$$

Feature Decomposition and Cascaded Integration (FDCI)

Inspired by [Gao *et al.*, 2019], we design an MS-Conv module to extract multi-scale information to obtain a larger receptive field without introducing too much overhead, thereby helping the FDCI obtain richer information.

Before the MS-Conv layer, we first employ a 1×1 convolutional layer to project f_{RoI} into a new feature $f_{local} \in \mathbb{R}^{\hat{c}\times h\times w}$. Subsequently, it is divided into m parts along channels, denoted $f_1, f_2, \cdots, f_m$.

$$f_1, f_2, \cdots, f_m = \text{Split}(\text{Conv}_{in}(f_{RoI})) \quad (5)$$

The process of MS-Conv can be described by Eq. 6. Given the i-th group features f_i, the output of the previous stage o_{i-1} is transformed by a 3×3 convolutional layer to get o'_{i-1}. Afterward, an element-wise weighted sum is adopted to calculate the i-th stage output o_i. A set of learnable parameters α_i and β_i are introduced as the aforementioned weights after sigmoid so that the appropriate summation ratio can be learned during optimization. To drive this recursion, we directly employ an identity projection as the first stage of MS-Conv, i.e. $o_1 = f_1$. Considering that each stage of MS-Conv is affected by previous stages, in the subsequent stage, the cascade operation makes the receptive field gradually increase. Thus MS-Conv makes the output feature include different scales feature representation, which further enhances the ability of FDCI to fuse local context information.

$$o_i = \sigma(\alpha_i) \cdot f_i + \sigma(\beta_i) \cdot \text{Conv}_i(o_{i-1}) \quad (6)$$

Subsequently, concatenate the m groups output of MS-Conv and feed it into a 1×1 convolutional layer to obtain the final output f_{FDCI}. That is, the output can be expressed as

$$FDCI(f_{RoI}) = \text{Conv}_{out}(\text{Concat}(o_1, o_2, \cdots, o_m)) \quad (7)$$

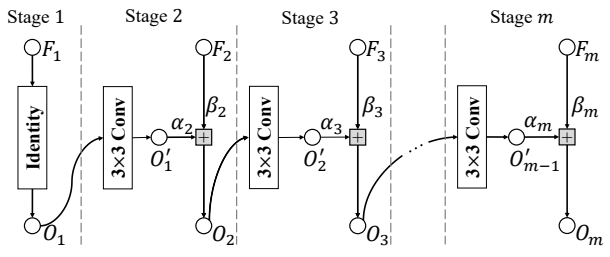


Figure 4: Our proposed MS-Conv

3.3 Loss Function

Just like [Xiao *et al.*, 2017], in this paper, we optimize our model with the following loss

$$\mathcal{L} = \mathcal{L}_{rpn} + \mathcal{L}_{reid} + \mathcal{L}_{cls} + \mathcal{L}_{box} \tag{8}$$

where $\mathcal{L}_{rpn}, \mathcal{L}_{reid}, \mathcal{L}_{cls}, \mathcal{L}_{box}$ represents the mainstream rpn loss, OIM loss, cross entropy loss, smooth-L1 loss, respectively.

4 Experiments

4.1 Datasets and Settings

Datasets

PRW. PRW contains 11816 surveillance video frames captured by 6 cameras deployed at Tsinghua University, including 43110 pedestrians with 932 different identities. The training set includes 5704 frames, of which there are 15575 pedestrians with 482 identities. The test set contains 450 different identities in 6112 frames and 2057 query probes in total. PRW performs retrieval using all gallery sets.

CUHK-SYSU. CUHK-SYSU is another mainstream person search dataset consisting of 18184 street scene photos and movie stills. The training set contains 11206 images and 5532 kinds of pedestrian identities, while the test set uses 2900 queries to search in 6978 gallery images. Only partially defined gallery subsets with various sizes are used for each query in the search process.

Evaluation Metrics

We adopt two evaluation methods, Cumulative Match Characteristic (CMC) and mean Average Precision (mAP), which are usually used in person search.

Implementation Details

We implement our method using PyTorch, and all experiments are performed on an NVIDIA GeForce RTX 3090. As mentioned before, we apply an off-the-shelf ResNet-50 pretrained on ImageNet as the backbone. In the MLFF, both MADE and FDCI output 512-channels feature representation, concatenated into 1024 channels to meet the requirement of the subsequent conv5_x. MADE uses an MHSA with 4 heads, for each head, the dimensions of key and query are set to 64 when the value is 128. Considering that the input features are only 14×7 in size, the stage of FDCI is set to 4, as a result, the receptive field of the last stage is just 7. During training, we employ 4 images resized to 1500×900 as a mini-batch. Our stem network and detection network are optimized by SGD, whose momentum is 0.9, and the initial learning rate is 0.0024. The identification network is Optimized by AdamW with an initial learning rate of 0.00028. For PRW and CUHK-SYSU, the length of the circular queue is set to 5000 and 500, respectively. Besides, random horizontal flip is employed for data augmentation, and tricks such as CWS and PKSample are also adopted.

4.2 Comparison with the State-of-the-art Methods

Comparison on PRW. Benefiting from the VSRI network, our method could extract identity-specific representation that only focuses on the uniqueness of pedestrian. The introduction of MLFF enables the re-ID network to pay more attention to the details of pedestrians and to obtain more discriminative feature representations. In Table 1, we compare our method with other state-of-the-art methods. The results show that our performance is much higher than other one-stage methods on the PRW dataset in terms of mAP and top-1. Even with two-stage methods such as IGPN and TCTS, our method still suppresses them by 5.7% and 6.1% on mAP. Applying ground truth instead of coarse proposal yields an additional 1.8% mAP and top-1 improvement, which further demonstrates the representational power of our method. Compared with CUHK-SYSU, PRW has less data, so it is more challenging. Such a significant improvement on PRW shows that our method is capable of extracting more representative features, which further proves the effectiveness of our methods.

Method	PRW		CUHK-SYSU	
	mAP	top-1	mAP	top-1
MGTS[Chen *et al.*, 2018]	32.6	72.1	83.0	83.7
CLSA[Lan *et al.*, 2018]	38.7	65.0	87.2	88.5
IGPN[Dong *et al.*, 2020b]	**47.2**	87.0	90.3	91.4
RDLR[Han *et al.*, 2019]	42.9	70.2	93.0	94.2
TCTS[Wang *et al.*, 2020]	46.8	**87.5**	**93.9**	**95.1**
OIM[Xiao *et al.*, 2017]	21.3	49.9	75.5	78.7
IAN[Xiao *et al.*, 2019]	23.0	61.9	76.3	80.1
NPSM[Liu *et al.*, 2017]	24.2	53.1	77.9	81.2
RCAA[Chang *et al.*, 2018]	-	-	79.3	81.3
CTXGraph[Yan *et al.*, 2019]	33.4	74.6	84.1	86.5
H-OIM[Chen *et al.*, 2020a]	39.8	80.4	89.7	90.8
BI-Net[Dong *et al.*, 2020a]	45.3	81.7	90.0	90.7
NAE[Chen *et al.*, 2020b]	43.3	80.9	91.5	92.4
NAE+[Chen *et al.*, 2020b]	44.0	81.1	92.1	92.9
DKD[Zhang *et al.*, 2021]	50.5	87.1	93.1	94.2
DMR-Net[Han *et al.*, 2021]	46.9	83.3	93.2	**94.2**
Ours	**52.9**	**87.3**	**93.4**	94.1
Ours + Ground Truth	**54.7**	**89.1**	**94.1**	**94.7**

Table 1: Comparison of mAP and top-1 with the state-of-the-art methods on PRW and CUHK-SYSU. The upper half belongs to two-stage methods, and the bottom half belongs to one-stage methods.

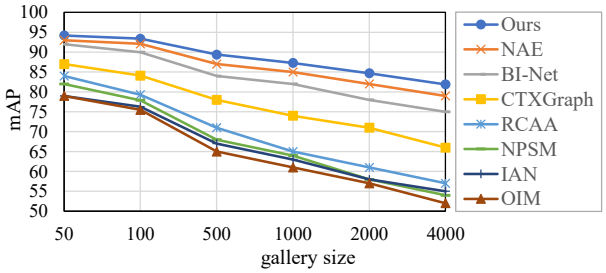


Figure 5: Comparison of mAP under different gallery sizes.

Comparison on CUHK-SYSU. Although CUHK-SYSU has more complex scenes, our MLFF design facilitates the model to focus more on the pedestrian itself without being disturbed by background information. In addition, the VSRI network isolates the representations of the two subtasks from

Method	re-ID		All Detection	
	mAP	top-1	Recall	AP
OIM	36.8	77.1	93.9	91.3
VSRI Network(GAP)	40.5	81.3	94.7	92.2
VSRI Network(GMP)	**41.3**	**81.4**	**94.8**	**92.3**

Table 2: Evaluation of the improvement of the VSRI on PRW.

each other, and each subtask can be optimized without interference, which makes us achieve better results on CUHK-SYSU as well. As demonstrated in Table 1, compared with other state-of-the-art methods, our method achieves the best mAP 93.2% and the second-best top-1 94.2% beyond the one-stage methods. Compared with most of the two-stage methods, our method still has advantages. We can also draw a conclusion that our VSRI network is also beneficial for the detection task to obtain more general detection capabilities, and is more conducive to the detection of hard samples. We find that [Han *et al.*, 2021] achieves better detection results with a high-accuracy detector RepPoints, although its top-1 accuracy is higher than our method, its mAP is lower, which explains the isolated detection branch could detect hard samples that are difficult to distinguish, resulting in higher mAP with improved recall. We also evaluate the scalability of our method and other state-of-the-art methods under different gallery sizes. As Shown in Figure 5, our method performs favorably against other one-stage methods among all gallery sizes.

5 Ablation Study

5.1 Evalutions of the Isolated Representation

The vision shared and representation isolated design of our proposed VSRI network enables detection subtask and re-ID subtask to obtain isolated task-specific representations respectively, thus solving the problem of the traditional shared representation method, which is difficult to simultaneously represent the consistency and discrimination of pedestrians in the shared features. Compared with the mainstream GAP, the application of GMP in the re-ID branch avoids the tendency of all human representations to be averaged and retains more differentiated pedestrian features. Experimental results summarized in Table 2 prove these advantages of our proposed VSRI network. We compare the VSRI network with the traditional shared representation structure, the recall and AP of the VSRI network are 0.9% and 1.0% higher than that of OIM, respectively, and the mAP and top-1 of re-ID subtask are increased by 4.5% and 4.3%. And without the influence of GMP, the detection AP and recall decrease by 0.1%, the performance of mAP and top-1 drop by 0.8% and 0.1%, respectively. This proves that our VSRI network enables both branches to simultaneously obtain task-specific feature representations, resulting in improved performance in both detection and re-ID.

5.2 Evalutions of the Sub-modules in MLFF

As mentioned earlier, the MLFF includes three parts: the SPFF, the MADE, and the FDCI. The SPFF establishes a spatial feature pyramid for each RoI so that its visual features could have richer details. The application of the MADE and the FDCI facilitates the MLFF to extract features of multiple scales from single-pixel to global context from the pyramidal features, and fully integrate them. Due to the complementarity of the three components, the MLFF can obtain fine-grained representation with more identity discrimination ability. The experimental results show that the fine-grained representation significantly improves the re-ID performance.

SPFF	MADE	FDCI	mAP	top-1
-	-	-	41.3	81.4
✓	-	-	51.2(↑ 9.9)	86.8(↑ 5.4)
✓	✓	-	52.4(↑ 11.1)	87.2(↑ 5.8)
✓	-	✓	51.7(↑ 10.4)	86.4(↑ 5.0)
✓	✓	✓	**52.9**(↑ 11.6)	**87.3**(↑ 5.9)

Table 3: Evaluation of the influence of the SPFF, the MADE, and the FDCI in the MLFF with different settings.

Table 3 records the results of our experiments. We use the aforementioned VSRI network as a baseline and verify the performance of several combinations of the three parts. During experiments, the unused MADE or FDCI are replaced with a 1×1 convolutional layer. The absence of SPFF means that only F_{C4} is used. Results demonstrate that introducing multi-level features increases mAP from 41.3% to 51.2%, and top-1 increased from 81.4% to 86.8% as well. With the addition of MADE, mAP and top-1 increased by 1.2% and 0.4%, respectively. Adding FDCI to SPFF, mAP could also increase by 0.5%. By combining these three parts, the mAP is finally increased to 52.9%, and the top-1 further achieves an excellent performance of 87.3%. We can conclude that the re-ID subtask relies more on fine-grained features than the detection subtask, and the fully fused feature provided by MLFF is vitally important for re-ID.

6 Conclusion

In this paper, we propose a Vision Shared and Representation Isolated (VSRI) network to obtain isolated task-specific representations to alleviate the natural conflicts between pedestrian detection and person re-ID. In addition, we design a multi-level feature fusion module, containing the SPFF, the MADE, and the FDCI, to obtain pyramidal visual features and fuse them in multiple scales from single-pixel to global. Extensive experiments have proved that our method significantly improves the performance compared to previous representation shared one-stage methods. This demonstrates that the person search models can benefit from decoupling two subtasks and introducing more fine-grained information.

Acknowledgments

This research is supported in part by the grants from the National Natural Science Foundation of China (No. 62006037, No. 61972074, No. 62172073), and in part by the National Key Research and Development Program of China (No. 2021YFB3301904).

References

[Chang *et al.*, 2018] Xiaojun Chang, Po-Yao Huang, Yi-Dong Shen, Xiaodan Liang, Yi Yang, and Alexander G Hauptmann. Rcaa: Relational context-aware agents for person search. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 84–100, 2018.

[Chen *et al.*, 2018] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Ying Tai. Person search via a mask-guided two-stream cnn model. In *Proceedings of the european conference on computer vision (ECCV)*, pages 734–750, 2018.

[Chen *et al.*, 2020a] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Bernt Schiele. Hierarchical online instance matching for person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10518–10525, 2020.

[Chen *et al.*, 2020b] Di Chen, Shanshan Zhang, Jian Yang, and Bernt Schiele. Norm-aware embedding for efficient person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12615–12624, 2020.

[Dong *et al.*, 2020a] Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Bi-directional interaction network for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2839–2848, 2020.

[Dong *et al.*, 2020b] Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Instance guided proposal network for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2585–2594, 2020.

[Gao *et al.*, 2019] Shanghua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip HS Torr. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[Han *et al.*, 2019] Chuchu Han, Jiacheng Ye, Yunshan Zhong, Xin Tan, Chi Zhang, Changxin Gao, and Nong Sang. Re-id driven localization refinement for person search. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9814–9823, 2019.

[Han *et al.*, 2021] Chuchu Han, Zhedong Zheng, Changxin Gao, Nong Sang, and Yi Yang. Decoupled and memory-reinforced networks: Towards effective feature learning for one-step person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1505–1512, 2021.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Lan *et al.*, 2018] Xu Lan, Xiatian Zhu, and Shaogang Gong. Person search by multi-scale matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 536–552, 2018.

[Liu *et al.*, 2017] Hao Liu, Jiashi Feng, Zequn Jie, Karlekar Jayashree, Bo Zhao, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. Neural person search machines. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 493–501, 2017.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[Wang *et al.*, 2020] Cheng Wang, Bingpeng Ma, Hong Chang, Shiguang Shan, and Xilin Chen. Tcts: A task-consistent two-stage framework for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11952–11961, 2020.

[Xiao *et al.*, 2016] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. End-to-end deep learning for person search. *arXiv preprint arXiv:1604.01850*, 2(2):4, 2016.

[Xiao *et al.*, 2017] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3415–3424, 2017.

[Xiao *et al.*, 2019] Jimin Xiao, Yanchun Xie, Tammam Tillo, Kaizhu Huang, Yunchao Wei, and Jiashi Feng. Ian: the individual aggregation network for person search. *Pattern Recognition*, 87:332–340, 2019.

[Xu *et al.*, 2014] Yuanlu Xu, Bingpeng Ma, Rui Huang, and Liang Lin. Person search in a scene by jointly modeling people commonness and person uniqueness. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 937–940, 2014.

[Yan *et al.*, 2019] Yichao Yan, Qiang Zhang, Bingbing Ni, Wendong Zhang, Minghao Xu, and Xiaokang Yang. Learning context graph for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2158–2167, 2019.

[Zhang *et al.*, 2021] Xinyu Zhang, Xinlong Wang, Jia-Wang Bian, Chunhua Shen, and Mingyu You. Diverse knowledge distillation for end-to-end person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3412–3420, 2021.

[Zheng *et al.*, 2017] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1367–1376, 2017.