MSINet: Twins Contrastive Search of Multi-Scale Interaction for Object ReID

Jianyang Gu1  Kai Wang2  Hao Luo3  Chen Chen4  Wei Jiang1∗
Yuqiang Fang5  Shanghang Zhang6  Yang You2  Jian Zhao7∗
1Zhejiang University  2National University of Singapore  3Alibaba Group
4OPPO Research Institute  5Space Engineering University  6Peking University
7Institute of North Electronic Equipment
{gu.jianyang, jiangwei_zju}@zju.edu.cn zhaojian90@u.nus.edu

Abstract

Neural Architecture Search (NAS) has been increasingly appealing to the society of object Re-Identification (ReID), for that task-specific architectures significantly improve the retrieval performance. Previous works explore new optimizing targets and search spaces for NAS ReID, yet they neglect the difference of training schemes between image classification and ReID. In this work, we propose a novel Twins Contrastive Mechanism (TCM) to provide more appropriate supervision for ReID architecture search. TCM reduces the category overlaps between the training and validation data, and assists NAS in simulating real-world ReID training schemes. We then design a Multi-Scale Interaction (MSI) search space to search for rational interaction operations between multi-scale features. In addition, we introduce a Spatial Alignment Module (SAM) to further enhance the attention consistency confronted with images from different sources. Under the proposed NAS scheme, a specific architecture is automatically searched, named as MSINet. Extensive experiments demonstrate that our method surpasses state-of-the-art ReID methods on both in-domain and cross-domain scenarios. Source code available in https://github.com/vimar-gu/MSINet.

1. Introduction

Object re-identification (Re-ID) aims at retrieving specific object instances across different views [39, 40, 57, 65, 70], which attracts much attention in computer vision community due to its wide-range applications. Previous works have achieved great progress on both supervised [42,49,58] and unsupervised ReID tasks [17,50,76], most of which adopts backbone models originally designed for general image classification tasks [20,52].

Recent literature [64,74] has shown that applying different architectures on ReID leads to large performance variations. Some works employ Neural Architecture Search (NAS) for ReID [28,45]. The proposed optimizing targets and search spaces stably improve the model performance, yet the main search scheme still follows traditional NAS methods designed for general classification tasks [12,36]. As an open-set task, ReID contains different categories in the training and validation sets [64,71], while the two sets share exactly the same categories in standard classification tasks [10], which is also followed by traditional NAS methods. The incompatibility between search schemes and real-world training schemes makes the searched architecture sub-optimal for ReID. Moreover, ReID is required to distinguish more subtle distinctions among fine-grained instances compared with image-level classification [48,63]. Some previous works [4,44,68,74] have manifested that local perspectives and multi-scale features are discriminative for ReID. However, current utilizations of these features are mostly empirically designed, which can be more flexible according to the characteristics of different network layers.

Figure 1. The left panel shows the example activation maps of ResNet50 (1st row) and MSINet (2nd row). The right panel shows the average distances between the most similar 10 negative samples and each query image at the inference. Best viewed in color.

*Co-corresponding authors
In this work, we propose a novel NAS scheme aiming at addressing the aforementioned challenges. In order to simulate the real-world ReID training schemes, a Twins Contrastive Mechanism (TCM) is proposed to unbind the categories of the training and validation sets. An adjustable overlap ratio of categories builds up the compatibility between NAS and ReID, which provides more appropriate supervision for ReID architecture search. Moreover, to search for more rational utilizations of multi-scale features, we design a Multi-Scale Interaction (MSI) search space. The MSI space focuses on interaction operations between multi-scale features along the shallow and deep layers of the network, which guides the features to promote each other. Additionally, to further improve the generalization capability, we propose a Spatial Alignment Module (SAM) to enhance the attention consistency of the model confronted with images from different sources. With the above NAS scheme, we obtain a light-weight yet effective model architecture, denoted as Multi-Scale Interaction Net (MSINet).

We visualize the example activation maps of our proposed MSINet and ResNet50 [20] trained on VeRi-776 [38, 39] in Fig. 1. Compared to ResNet50, MSINet focuses on more unique distinctions with specific semantic information to recognize instances. Besides, MSINet largely increases the distance margin between query image and corresponding negative samples, reflecting extraordinary discriminative capability. Extensive experiments demonstrate that MSINet surpasses state-of-the-art (SOTA) ReID methods on both in-domain and cross-domain scenarios. Our source codes are available in the supplementary material.

Our contributions are summarized as follows:

- To the best of our knowledge, we are the first to build the NAS search scheme according to the real-world ReID training schemes, which provides more appropriate supervision for the ReID architecture search.
- We propose a novel search space based on the Multi-Scale Interaction (MSI) operations and a Spatial Alignment Module (SAM) to improve the model performance on in-domain and cross-domain scenarios.
- We construct a light-weight yet effective architecture for ReID tasks, denoted as MSINet. With only 2.3M parameters, MSINet surpasses ResNet50 [20] by 9% mAP on MSMT17 [60] and 16% mAP on MSMT17 → Market-1501 [69].

2. Related Works

Neural Architecture Search. NAS has been increasingly appealing to the computer vision society, due to its automatic architecture designing characteristics. NAS methods can be roughly separated into four categories: reinforcement learning [1, 75], evolutionary algorithms [35, 46], gradient descent [36, 43] and performance prediction [11, 34]. Liu et al. establish a differentiable architecture search (DARTS) method [36], which improves the practicability of NAS by a large extent. Some later works further improve the structure through sampling strategy [62], network pruning [2, 8], progressive learning [5], collaborative competition [7], etc. Most of NAS works focus on general image classification tasks, where the training and validation sets share the exact same categories. Following the setting, however, leads to incompatibility with the real-world training schemes of object ReID. In this work, we unbind the category bond between the two sets and propose a novel search scheme suitable for ReID.

ReID Network Design. Current ReID works mostly adopt backbones designed for image classification [20, 23, 47, 52]. Some works [14, 31, 67] design attention modules based on the common backbones to unearth their potential on distinguishing local distinctions. However, these methods usually lead to high calculation consumption.

There are also several works focusing on designing ReID-specific architectures. Li et al. present a Filter Pairing Neural Network to dynamically match patches in the feature maps [30]. Wang et al. separate and regroup the features of two samples with a WConv layer [56]. Guo et al. extract multi-scale features to directly evaluate the similarity between samples [18]. However, the siamese structure is inconvenient when conducting retrieval on large galleries. Zhou et al. aggregate multi-scale information to achieve high accuracy with small computing consumption [74]. Quan et al. introduce a part-aware module into the DARTS search space [36, 45]. Li et al. propose a new search space in regard to receptive field scales [28]. These methods have excellent performance on limited parameter scales, but fail to surpass those networks with complex structures. Different from previous works, we design a light-weight searching structure focusing on rational interaction operations between multi-scale features. The searched MSINet surpasses SOTA methods on both in-domain and cross-domain tasks.

3. Methods

Our goal is to construct an effective NAS scheme to search for a light-weight backbone architecture suitable for ReID tasks. Based on the training schemes of ReID, we propose a novel Twins Contrastive Mechanism to provide more appropriate supervision for the search process. Aiming at rational interaction between multi-scale features, we design a Multi-Scale Interaction search space. We further introduce a Spatial Alignment Module to improve the generalization capability with limited parameter growth.

3.1. Twins Contrastive Mechanism

NAS aims at automatically searching for the optimal network architecture for certain data. Inspired by [36], a basic
differentiable architecture search scheme is established. We define the ordinary model parameters as $\omega$, and architecture parameters as $\alpha$. For network layer $i$ with a search space of $O$, $\alpha_i$ controls the weight of each operation $o$ in the space. The features are parallelly passed through all the operations, and the final output is formulated by the softmax-weighted sum of operation outputs:

$$f(x_i) = \sum_{o \in O} \frac{\exp \{\alpha_i^o\}}{\sum_{o' \in O} \exp \{\alpha_i^{o'}\}} \cdot o(x_i).$$ (1)

The search process is conducted in an alternative manner. Training data is utilized to update the model parameters, and validation data is then employed to update the architecture parameters. For most NAS methods designed for image classification tasks, the training and validation data share exactly the same categories and a linear classification layer for loss calculation.

![Figure 2](image.png)

Figure 2. The model structure of the proposed MSINet. The input can be either person or vehicle. Inside a cell, the input is separated to two branches, with different receptive field scales. The interaction module exchanges information between two branches. Architecture search automatically select the most appropriate interaction for each cell.

As the loss calculation does not rely on the linear classification layer, the categories of the training and validation sets are unbound. We are able to dynamically adjust the category overlap ratio in these two sets. The advantages of a proper overlap ratio are summarized as two folds. Firstly, TCM better simulates the real-world training of ReID and helps the model focus on truly discriminative distinctions. The differences between the training and validation data improve the generalization capability of the model. Secondly, a relatively small proportion of overlapped categories stabilizes the architecture parameter update through a consistent optimizing target with the model parameter update.

### 3.2. Multi-Scale Interaction Space

Although the local perspective and multi-scale features have already been investigated in previous ReID works [4, 28, 44, 51, 68, 74], the utilization of these information is mainly empirically designed aggregation, which is monotonous and restrained. We argue that on the one hand, the rational utilization of multi-scale features should be dynamically adjusted along the shallow and deep layers of the network. On the other hand, introducing interaction other than aggregation creates direct information exchange,
and makes fuller use of multi-scale features. Therefore, we propose a novel Multi-Scale Interaction (MSI) search space to establish a light-weight architecture suitable for ReID.

As shown in Fig. 2, the network is mainly grouped with MSI cells and down-sample blocks, which is generally consistent with OSNet [74]. In each cell, the input features are passed through two branches with different receptive field scales. To reduce the calculation burden of the network, for the layers inside each branch, we adopt the stack of 1×1 convolution and multiple depth-wise 3×3 convolution to implement specific scales. A scale ratio \( \rho \) of 3:1 is selected for the two branches. These two branches do not share model parameters, except for the Interaction Modules (IM). IM introduces information exchange for the two branches. There are 4 operation options for the IM. With the two-branch input features defined as \( (x_1, x_2) \), the operations can be formulated as:

**None.** None operation involves no parameters, and outputs exactly the input features \( (x_1, x_2) \).

**Exchange.** Exchange acts as the strongest interaction among all options. It directly exchanges the features for the two branches and outputs \( (x_2, x_1) \). Exchange contains no extra parameters, as well.

**Channel Gate.** Channel gate introduces a Multi-Layer Perceptron (MLP) to generate a channel-wise attention gate [61,74] as:

\[
G(x) = \sigma(MLP(x)),
\]

and returns \( (G(x_1), x_1, G(x_2) \cdot x_2) \). The MLP is composed of 2 fully connected layers and its parameters are shared for both branches. Thereby it achieves interaction by jointly screening discriminative feature channels.

**Cross Attention.** Traditional channel attention module calculates the channel correlation inside a single feature map [15]. The original feature map \( x \in \mathbb{R}^{C \times H \times W} \) is firstly reshaped into the query feature \( \tilde{x} \in \mathbb{R}^{C \times N} \), where \( N = H \times W \). Then the correlation activation is calculated by performing a matrix multiplication between the query feature \( \tilde{x} \) and the key feature \( \tilde{x}^\top \). We propose to exchange the keys of the two branches to explicitly calculate the correlation between each other. The correlation activation is then transformed to a mask, and is added up to the original features with a learnable proportion.

After interaction, the multi-scale branches are fused through a sum operation. It is worth noting that the extra parameters brought by multiple interaction options are limited, which enables searching for each cell along the whole network independently. At the beginning of the network, we employ the same stem module as that in OSNet [74], containing a 7 × 7 convolutional layer and a 3 × 3 max pooling with a stride of 2. After the searching process, the interaction operation \( o \) with the largest weight \( o^*_\alpha \) at each layer is reserved to form the searched architecture.

After searching the architecture, the model is validated on various Re-ID tasks. The training is constrained by the

---

Table 1. The detailed interaction operation in the proposed MSINet architecture. N: None; E: Exchange; G: Channel Gate; A: Cross Attention.

<table>
<thead>
<tr>
<th>Cell #1</th>
<th>Cell #2</th>
<th>Cell #3</th>
<th>Cell #4</th>
<th>Cell #5</th>
<th>Cell #6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>G</td>
<td>G</td>
<td>E</td>
<td>G</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>N</td>
<td>G</td>
<td>A</td>
<td>E</td>
<td>A</td>
<td></td>
</tr>
</tbody>
</table>

---

Figure 3. The aligning pipeline of the proposed Spatial Alignment Module. The correlation activation vectors are calculated between the anchor feature and all features in a mini-batch. The positive vectors are aligned with the learnable self-activation, and the negative vectors are aligned with each other. The structure of PAM is shown on the left.

Figure 4. Output feature maps of MSI cells at different layers.
We evaluate the consistency between activation vectors with the above process is denoted as “Mutual Conv” in Fig. 3. The retrieval precision of object ReID tasks are largely affected by the variation of appearances such as poses, illumination and occlusion when the camera conditions change. In order that the model correctly and consistently focuses can be many different hints for recognition, some of which might be inappropriate, such as the backgrounds. By aligning all the correlations for sample \( i \), we hope that the network can correct some attention bias and consistently focus on discriminative positions.

However, through aligning positive sample pairs, the ID-related features are expected to be emphasized, which cannot be achieved by aligning negative pairs. Therefore, we introduce an extra position activation module (PAM) to generate supervision for the alignment between positive pairs. The spatial alignment loss is formulated as:

\[
\mathcal{L}_{sa}(i) = \frac{1}{N_+} \sum \limits_{p \in \mathcal{I}_+} \left( 1 - S(\hat{a}(i), a(i, p)) \right) + \frac{1}{N_+} \sum \limits_{n_1, n_2 \in \mathcal{I}_-} \left( 1 - S(a(i, n_1), a(i, n_2)) \right),
\]

where \( \mathcal{I}_+ \) contains positive indices for sample \( i \), the total number of which is \( N_+ \), and \( \mathcal{I}_- \) contains negative indices for sample \( i \). For simplicity, we take the maximum activation for each position of sample \( i \) as:

\[
a(i, j) = \max \limits_{dwm=1} A(i, j).
\]

Specifically, we first calculate the position-wise correlation activation map \( A \) between the feature maps in a mini-batch. The activation between sample \( i \) and \( j \) can be formulated as: \( A(i, j) = x_j^T \times \hat{x}_i \), where \( \hat{x}_i \in \mathbb{R}^{C \times N} \) is the reshaped from the original feature \( x \in \mathbb{R}^{C \times H \times W} \). Then we take the maximum activation for each position of sample \( i \) as:

\[
a(i, j) = \max \limits_{dwm=1} A(i, j).
\]

The above process is denoted as “Mutual Conv” in Fig. 3. We evaluate the consistency between activation vectors with cosine similarity. For negative samples specifically, there

### Table 2. Supervised performance on object ReID datasets. The results in the top part are trained from scratch, and those in the bottom part are pre-trained on ImageNet in advance. As the compared methods are originally proposed for person ReID, we reproduce the results in vehicle datasets. * indicates that the results of person ReID are reproduced by us. The evaluation results of architecture searched on VR can be found in the supplementary material.

<table>
<thead>
<tr>
<th>Method</th>
<th>Params</th>
<th>Inference</th>
<th>M</th>
<th>R-1↑</th>
<th>MS</th>
<th>R-1↑</th>
<th>VR</th>
<th>R-1↑</th>
<th>VID</th>
<th>R-1↑</th>
<th>MS→M</th>
<th>R-1↑</th>
<th>VR→VID</th>
<th>R-1↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet50* [42]</td>
<td>~24M</td>
<td>1×</td>
<td>85.7</td>
<td>68.3</td>
<td>48.0</td>
<td>25.7</td>
<td>92.8</td>
<td>69.9</td>
<td>70.6</td>
<td>76.6</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OSNet [74]</td>
<td>2.2M</td>
<td>0.79×</td>
<td>93.6</td>
<td>81.0</td>
<td>71.0</td>
<td>43.3</td>
<td>95.4</td>
<td>72.8</td>
<td>76.0</td>
<td>88.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CDNet [28]</td>
<td>1.8M</td>
<td>0.67×</td>
<td>93.7</td>
<td>83.7</td>
<td>73.7</td>
<td>48.5</td>
<td>94.3</td>
<td>73.0</td>
<td>74.5</td>
<td>88.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MSINet</td>
<td>2.3M</td>
<td>0.71×</td>
<td>94.6</td>
<td>87.0</td>
<td>76.0</td>
<td>52.5</td>
<td>95.9</td>
<td>75.0</td>
<td>76.5</td>
<td>89.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

ResNet50* [42] ~24M 1× 94.5 85.9 75.5 50.4 94.5 73.6 76.5 89.9 58.8 31.8 42.8 61.9
OSNet [74] 2.2M 0.79× 94.8 84.9 78.7 52.9 95.5 76.4 76.0 88.6 66.6 37.5 46.5 63.1
CDNet [28] 1.8M 0.67× 95.1 86.0 78.9 54.7 - - - - - - - -
MSINet 2.3M 0.71× 95.3 89.6 81.0 59.6 96.8 78.8 77.9 91.7 74.9 46.2 48.0 65.6
MSINet-SAM 2.4M 0.71× 95.5 89.9 80.7 59.5 96.7 79.0 78.0 91.9 76.3 48.4 49.0 66.8

classification id loss and the triplet loss, formulated by:

\[
\mathcal{L}_{id} = \frac{1}{N} \sum \limits_{i=1}^{N} - \log \left( \frac{\exp W^T_i f_i}{\sum_j \exp W^T_j f_j} \right),
\]

where \( f_i \) is a feature vector, the corresponding classifier weight of which is \( W_i \), and

\[
\mathcal{L}_{tri} = [D(f_{a}, f_{p}) - D(f_{a}, f_{n}) + m]_+,
\]

where \( f_a, f_p, f_n \) are the embedded features for the anchor, the hardest positive and negative samples in a mini-batch, \( D(\cdot, \cdot) \) is the Euclidean distance, \( m \) is the margin parameter, and \( [\cdot]_+ \) is the \( \max(\cdot, 0) \) function.

3.3. Spatial Alignment Module

The retrieval precision of object ReID tasks are largely affected by the variation of appearances such as poses, illumination and occlusion when the camera conditions change. In order that the model correctly and consistently focuses on the discriminative spatial positions, we design a Spatial Alignment Module (SAM) to explicitly align the spatial attention between images, as shown in Fig. 3.

Specifically, we first calculate the position-wise correlation activation map \( A \) between the feature maps in a mini-batch. The activation between sample \( i \) and \( j \) can be formulated as: \( A(i, j) = x_j^T \times \hat{x}_i \), where \( \hat{x}_i \in \mathbb{R}^{C \times N} \) is reshaped from the original feature \( x \in \mathbb{R}^{C \times H \times W} \). Then we take the maximum activation for each position of sample \( i \) as:

\[
a(i, j) = \max \limits_{dwm=1} A(i, j).
\]

The above process is denoted as “Mutual Conv” in Fig. 3. We evaluate the consistency between activation vectors with cosine similarity. For negative samples specifically, there

4. Experiments

4.1. Datasets and Evaluation Metrics

Our proposed method is evaluated on two person ReID datasets Market-1501 [69], MSMT17 [60], and two vehicle ReID datasets VeRi-776 [38, 39] and VehicleID [37]. For simplicity, the four datasets are denoted as M, MS, VR and VID in the following sections, respectively. Evaluation metrics include Cumulative Matching Characteristic (CMC) and mean average precision (mAP), which are commonly utilized on ReID tasks.
Table 3. Supervised performance comparison between MSINet and SOTA methods on M and MS datasets.

<table>
<thead>
<tr>
<th>Method</th>
<th>M</th>
<th></th>
<th></th>
<th></th>
<th>MS</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R-1↑</td>
<td>mAP↑</td>
<td>R-1↑</td>
<td>mAP↑</td>
<td></td>
<td>R-1↑</td>
<td>mAP↑</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCB [51]</td>
<td>93.8</td>
<td>81.6</td>
<td>68.2</td>
<td>40.4</td>
<td></td>
<td>81.6</td>
<td>68.2</td>
<td>40.4</td>
<td></td>
</tr>
<tr>
<td>MGN [55]</td>
<td>95.7</td>
<td>86.9</td>
<td>76.9</td>
<td>52.1</td>
<td></td>
<td>95.3</td>
<td>89.6</td>
<td>81.0</td>
<td>59.6</td>
</tr>
<tr>
<td>OSNet [74]</td>
<td>93.6</td>
<td>81.0</td>
<td>71.0</td>
<td>43.3</td>
<td></td>
<td>93.6</td>
<td>81.0</td>
<td>71.0</td>
<td>43.3</td>
</tr>
<tr>
<td>IANet [22]</td>
<td>94.4</td>
<td>83.1</td>
<td>75.5</td>
<td>46.8</td>
<td></td>
<td>94.4</td>
<td>83.1</td>
<td>75.5</td>
<td>46.8</td>
</tr>
<tr>
<td>DGNet [72]</td>
<td>94.8</td>
<td>86.0</td>
<td>77.2</td>
<td>52.3</td>
<td></td>
<td>94.8</td>
<td>86.0</td>
<td>77.2</td>
<td>52.3</td>
</tr>
<tr>
<td>Auto-ReID</td>
<td></td>
<td></td>
<td>95.1</td>
<td></td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAN [26]</td>
<td>96.1</td>
<td>88.0</td>
<td>79.2</td>
<td>55.7</td>
<td></td>
<td>96.1</td>
<td>88.0</td>
<td>79.2</td>
<td>55.7</td>
</tr>
<tr>
<td>CDNet [28]</td>
<td>95.1</td>
<td>86.0</td>
<td>78.9</td>
<td>54.7</td>
<td></td>
<td>95.1</td>
<td>86.0</td>
<td>78.9</td>
<td>54.7</td>
</tr>
<tr>
<td>BAT-Net [14]</td>
<td>95.1</td>
<td>87.4</td>
<td>79.5</td>
<td>56.8</td>
<td></td>
<td>95.1</td>
<td>87.4</td>
<td>79.5</td>
<td>56.8</td>
</tr>
<tr>
<td>SFT [41]</td>
<td>94.1</td>
<td>87.5</td>
<td>79.0</td>
<td>58.3</td>
<td></td>
<td>94.1</td>
<td>87.5</td>
<td>79.0</td>
<td>58.3</td>
</tr>
<tr>
<td>CTF [66]</td>
<td>94.8</td>
<td>87.7</td>
<td></td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RGA-SC [67]</td>
<td>96.1</td>
<td>88.4</td>
<td>80.3</td>
<td>57.5</td>
<td></td>
<td>96.1</td>
<td>88.4</td>
<td>80.3</td>
<td>57.5</td>
</tr>
<tr>
<td>MSINet</td>
<td>95.3</td>
<td>89.6</td>
<td>81.0</td>
<td>59.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.2. Architecture Search

We conduct the searching process on MSMT17. SGD is adopted for model parameter update with an initial learning rate of 0.025. The model is trained for 350 epochs in total. We adopt a warm-up strategy for the first 10 epochs. Then the learning rate is decayed by 0.1 at 150, 225 and 300 epochs, respectively. Adam [27] is adopted for the architecture parameter update with an initial learning rate of 0.002. The learning rate is decayed at the same pace. The images are reshaped to $256 \times 128$ for person and $256 \times 256$ for vehicles. Data augmentation includes random flip, random crop and random erasing [73]. The searched architecture is presented in Tab. 1. The “MSINet” in the following experiment sections refers to this architecture.

We visualize the feature maps extracted by each MSI cell in Fig. 4. At the shallow layers of the network, the kernels mainly focus on overall contour information. Channel gate helps to filter out inferior information, such as the background. As we approach deeper layers, the extracted features each have specific semantic information, where cross attention is more likely to be selected for the interaction. It indicates that cross attention is more rational for exchanging high-level semantic information.

4.3. Comparison with Other Backbones

We first compare our proposed MSINet with ResNet50 and recent proposed light-weight backbones in both in-domain and cross-domain ReID scenarios.

In-Domain Tasks. We adopt a two-group supervised evaluation scheme similar to that in [28, 74]: training from scratch and fine-tuning ImageNet [10] pre-trained models. The training parameters for both schemes are kept the same as that in architecture search, except for an initial learning rate of 0.065. Triplet loss and cross entropy loss are adopted for the parameter update. The margin $m$ in Eq. 6 is set as 0.3. [28] adopts an FBLNeck. We also employ the same structure. The results are shown in Tab. 2.

ResNet50 is the most commonly utilized backbone network in ReID tasks, yet holds the worst performance. Moreover, ResNet50 largely depends on ImageNet pre-training, while MSINet without pre-training has already surpassed pre-trained ResNet50 on all metrics. Compared with the other datasets, MS contains more variations on illumination, background and camera pose, and brings a large performance gap between ResNet50 and other methods. It also validates the inadequacy of image classification networks on ReID tasks. OSNet [74] and CDNet [28] are recently proposed architectures designed specifically for ReID tasks. Both architectures focus on fusing multi-scale features to better suit ReID. CDNet employs a traditional NAS scheme to search for the proper receptive field scales for each cell. MSINet fixes the receptive field scale and instead selects optimal interaction operations inside each cell. With only a bit more parameters, MSINet surpasses all the other backbones by a large margin.

Cross-Domain Tasks. Cross-domain experiments verify the generalization capability of the model. Following previous domain generalizable ReID works [25, 33], data augmentation is adjusted to random flip, random crop and color jittering. The model is pre-trained and fine-tuned for 250 epochs to avoid over-fitting. The other settings are kept the same as supervision scenes. With no present pre-trained models for CDNet [28], it is excluded from this section.

Tab. 2 shows that ResNet50 can be easily interfered by different image styles confronted with new image domains. OSNet learns multi-scale features with specific semantic information for ReID, which is domain invariant to some extent. Our proposed search scheme also takes into ac-
Figure 5. Ablation studies on (a) search scheme; (b) architecture; (c) multi-scale aggregation; (d) scale ratio $\rho$; (e) spatial alignment weight.

Table 5. The effectiveness of each component in SAM. Both in-domain performance on VR and cross-domain performance on VR→VID is evaluated.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MSINet</td>
<td>✓</td>
<td>✓</td>
<td>Self</td>
<td>96.7</td>
<td>78.4</td>
<td>48.5</td>
<td>66.2</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>Unified</td>
<td>96.6</td>
<td>78.3</td>
<td>48.3</td>
<td>65.8</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>Separated</td>
<td>96.8</td>
<td>78.6</td>
<td>48.7</td>
<td>66.4</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>PAM-Self</td>
<td>96.7</td>
<td>79.0</td>
<td>49.0</td>
<td>66.8</td>
</tr>
<tr>
<td>OSNet</td>
<td>✓</td>
<td>✓</td>
<td>None</td>
<td>95.5</td>
<td>76.4</td>
<td>46.5</td>
<td>63.1</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>PAM-Self</td>
<td>95.9</td>
<td>76.3</td>
<td>47.5</td>
<td>63.3</td>
</tr>
</tbody>
</table>

We count the generalization capability of the model. By partly separating the categories for training and validation sets, the searched interaction operations generalize well confronted with new image domains. Except for discrimination, MSINet also surpasses the other backbones on cross-domain tasks by a large margin with faster inference speed.

Additionally, we introduce SAM into the model, which aligns the spatial correlations between images. A weighted sum of ReID loss and spatial alignment loss is utilized when training the network with SAM. The weight of spatial alignment loss is set as $\lambda_{sa} = 2.0$. Without extra inference consumption or damages on the supervised performance, SAM further boosts the generalization capability of MSINet.

4.4. Comparison with State-of-the-art Methods

Tab. 3 further illustrates the supervision performance comparison of our proposed MSINet with the SOTA methods on M and MS datasets. With much less parameters than most of the compared methods, MSINet achieves a retrieval accuracy comparable to that of more complicated ones. Auto-ReID [45] first designs a NAS scheme for ReID, yet the DARTS-style architecture contains 13M parameters. RGA-SC [67] carefully designs a relation-aware global attention module. MSINet achieves even higher performance with less training consumption, which validates the superiority of selecting rational interaction.

We also evaluate the model performance replacing the backbone network from ResNet50 to MSINet for SOTA unsupervised ReID methods in Tab. 4. For purely unsupervised learning (USL) method GS [19] on M dataset, MSINet performs slightly lower on rank-1, yet has a large superiority on mAP. For HDCRL [6], MSINet shows obvious superiority over ResNet50. For unsupervised domain adaptation (UDA) method IDM [9] on M→MS task, MSINet surpasses ResNet50 by a large margin, which further proves that the TCM brings outstanding generalization capability to the searched architecture.

4.5. Ablation Studies

Effectiveness of Architecture Search. To verify the effectiveness, we conduct supervised training on MS with different search schemes in Fig. 5 (a). Under the standard classification scheme (“CE Overlap”), the searched model performs poor. Replacing the cross entropy loss to contrastive loss (“TCM Overlap”) only brings a slight improvement. The complete TCM framework unbounds the categories between the training and validation set, and thereby improves the performance by a large margin.

In Fig. 5 (b) we compare the performance of different architectures. Firstly, we validate 4 models each with a unique interaction operation from the 4 options in the search space. None and Exchange, with no trainable parameters, achieve poor performance. Channel gate introduces channel-wise attention, whose model performs the best among 4 options. Cross attention exchanges the key features for the two branches. Over-frequent exchange interferes the ordinary feature extraction and degrades the network performance. Through appropriately arranging interaction operations along the architecture, MSINet surpasses all the above 4 models. Random architecture, on the other hand, shows no rational appliance of interaction operations, which validates that the proposed search scheme helps find
### Table 6. Supervised performance comparison between MSINet and Transformers on VR and MS datasets.

<table>
<thead>
<tr>
<th>Method</th>
<th>Params</th>
<th>Inference Time</th>
<th>MS R-1↑ mAP↑</th>
<th>VR R-1↑ mAP↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeiT-S [21]</td>
<td>∼22M</td>
<td>0.97x</td>
<td>76.3</td>
<td>55.2</td>
</tr>
<tr>
<td>DeiT-B [21]</td>
<td>∼86M</td>
<td>1.79x</td>
<td>81.9</td>
<td>61.4</td>
</tr>
<tr>
<td>ViT-B [21]</td>
<td>∼86M</td>
<td>1.79x</td>
<td>81.8</td>
<td>61.0</td>
</tr>
<tr>
<td>MSINet</td>
<td>2.3M</td>
<td>0.71x</td>
<td>81.0</td>
<td>59.6</td>
</tr>
</tbody>
</table>

- **Figure 6.** Example top-15 retrieved sequences comparison on VR. Appearance differences caused by variant camera conditions are well addressed by the proposed MSINet. Visualization of person ReID can be found in the supplementary material.

- **Effectiveness of Spatial Alignment Module.** We validate the effectiveness of each components of SAM on the VR→VID cross-domain experiment in Tab. 5. Firstly, we introduce the spatial alignment for positive and negative sample pairs, respectively. Each of them brings certain performance improvements. However, a unified alignment for all sample pairs damages ID-related features and degrades the performance instead. Therefore, we separate the alignment of positive and negative samples, which retains some discriminative features and integrates the effect of both aspects. The extra PAM for positive sample alignment further guarantees the focus on ID-related positions and achieves the best performance. We also conduct in-domain experiment on VR to prove that SAM improves the generalization capability without sacrificing the supervision performance. Adding SAM to OSNet receives similar results, which validates the universality of SAM.

- **Fusing Operation.** After interaction, the multi-scale features are fused by sum operation. We investigate several fusing options on MS training from scratch in Fig. 5 (c). Subtracting (“Minus”) a branch from the other leads to the same results as “Sum” while multiplication (“Mul”) performs poorly.

- **Comparison with Transformer.** Transformer, as a new architecture, has recently been continuously making progresses in many computer vision domains [13,53], including ReID [21,32]. We compare the model performance with some baseline Transformer models in Tab. 6. DeiT-B and ViT-B [21] achieves higher performance on MSMT17, with much larger calculation burden compared with our proposed MSINet. On VeRi-776, MSINet surpasses all the baseline Transformer methods. It proves that rational interaction operations between multi-scale features are capable to assist light-weighted pure-CNN models to obtain comparable performance with complex Transformers.

- **Parameter Analysis.** Firstly, we study the influence of different receptive field scale ratios $\rho$ inside an MSI cell on MS training from scratch in Fig. 5 (d). Introducing scale differences between branches improves the model performance significantly, and subsequent increases brings more modest impacts. Considering both parameter scales and model performance, the ratio of 3:1 is selected for MSINet.

- **Visualization Results.** We visualize the top-15 retrieved sequences and the corresponding distances from an query image on VR in Fig. 6. By comparison, ResNet50 mainly focuses on general appearance features, where the top-rated negative samples share similar car bodies. MSINet, oppositely, concentrates on discriminative distinctions, an empty car hopper in this case, and creates an evident distance gap between positive and negative samples. More details can be seen in the supplementation material.

## 5. Conclusion

In this paper, we design a Twins Contrastive Mechanism for NAS to build the compatibility with ReID. The task-specific search scheme provides the searching process with more appropriate supervision. A Multi-Scale Interaction search space is proposed to establish rational and flexible utilization of multi-scale features. With a Spatial Alignment Module, our proposed MSINet achieves SOTA performance on both supervision and cross-domain scenarios with limited parameter amount. We hope the proposed approach could inspire more works focusing on designing network architectures suitable for ReID tasks.

## 6. Acknowledgement

This work is partially supported by National Natural Science Foundation of China (62173302, 62006244), China Scholarship Council (202206320302), Young Elite Scientist Sponsorship Program of China Association for Science and Technology (YESS20200140), and AI Singapore Programme (AISG2-PhD-2021-08-008).
References


[34] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathan Shlens, Wei Hu, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. In ECCV, pages 19–34, 2018. 2


[38] Xinchen Liu, Wu Liu, Huadong Ma, and Huiyuan Fu. Large-scale vehicle re-identification in urban surveillance videos. In ICME, pages 1–6, 2016. 2, 5


[57] Zhikang Wang, Lihuo He, Xiaoguang Tu, Jian Zhao, Xinbo Gao, Shengmei Shen, and Jiashi Feng. Robust video-based person re-identification by hierarchical mining. T-CSVT, 2021. 1


[63] Cheng Yan, Guansong Pang, Xiao Bai, Changhong Liu, Ning Xin, Lin Gu, and Jun Zhou. Beyond triplet loss: person re-identification with fine-grained difference-aware pairwise loss. T-MM, 2021. 1

Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Deep metric learning for person re-identification. In ICPR, pages 34–39, 2014. 1


