# Counterfactual Intervention Feature Transfer for Visible-Infrared Person Re-identification

Xulin Li[1,2*], Yan Lu[1,2*], Bin Liu[1,2✉], Yating Liu[3], Guojun Yin[1,2], Qi Chu[1,2], Jinyang Huang[1,2], Feng Zhu[4], Rui Zhao[4,5], and Nenghai Yu[1,2]

[1] School of Information Science and Technology, University of Science and Technology of China
[2] Key Laboratory of Electromagnetic Space Information, Chinese Academy of Science
[3] School of Data Science, University of Science and Technology of China
[4] SenseTime Research
[5] Qing Yuan Research Institute, Shanghai Jiao Tong University
{lxlkw,luyan17}@mail.ustc.edu.cn, flowice@ustc.edu.cn
{liuyat,gjyin}@mail.ustc.edu.cn, qchu@ustc.edu.cn
huangjy@mail.ustc.edu.cn, {zhufeng,zhaorui}@sensetime.com
ynh@ustc.edu.cn

**Abstract.** Graph-based models have achieved great success in person re-identification tasks recently, which compute the graph topology structure (affinities) among different people first and then pass the information across them to achieve stronger features. But we find existing graph-based methods in the visible-infrared person re-identification task (VI-ReID) suffer from bad generalization because of two issues: 1) **train-test modality balance gap**, which is a property of VI-ReID task. The number of two modalities data are balanced in the training stage but extremely unbalanced in inference, causing the low generalization of graph-based VI-ReID methods. 2) **sub-optimal topology structure** caused by the end-to-end learning manner to the graph module. We analyze that the joint learning of backbone features and graph features weaken the learning of graph topology, making it not generalized enough during the inference process. In this paper, we propose a Counterfactual Intervention Feature Transfer (CIFT) method to tackle these problems. Specifically, a Homogeneous and Heterogeneous Feature Transfer ($H^2FT$) is designed to reduce the train-test modality balance gap by two independent types of well-designed graph modules and an unbalanced scenario simulation. Besides, a Counterfactual Relation Intervention (CRI) is proposed to utilize the counterfactual intervention and causal effect tools to highlight the role of topology structure in the whole training process, which makes the graph topology structure more reliable. Extensive experiments on standard VI-ReID benchmarks demonstrate that CIFT outperforms the state-of-the-art methods under various settings.

**Keywords:** Person Re-identification, Counterfactual, Cross-modality

[*] Equal contribution.
[✉] Corresponding authors.

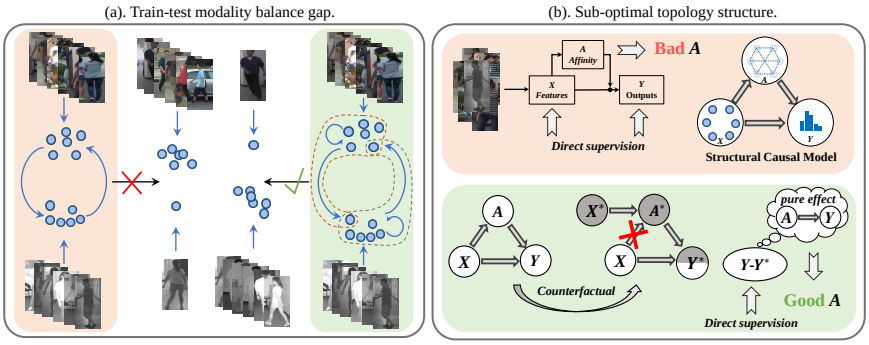


**Fig. 1.** The red background hints at the existing methods and the green one means our method. (a) Existing graph-based methods trained on modality-balanced data are difficult to transfer to a modality unbalanced scenario. Our method overcomes this problem by unbalanced scenarios simulation and a novel graph module design. (b) Existing training strategies learn the backbone features and the graph outputs jointly, making affinity learning be weakened. Our method uses a counterfactual intervention tool to calculate the pure effect contributed by the affinity changes only, making the model perceive the role of graph topology more direly.

## 1 Introduction

Standard person re-identification (ReID) [5, 15, 25, 26, 43] aims to match pedestrian images of the same identity captured by different cameras, which is essentially a single-modality (RGB) retrieval task. However, open-world intelligent monitoring requires methods to retrieve targets captured by infrared or thermal cameras in the dark scenario. Therefore, the research on visible-infrared person re-identification (VI-ReID) has attracted great attention in recent years. Different from standard ReID, large cross-modality discrepancy and intra-modality variations bring new challenges to research on VI-ReID. Most researchers [7, 11, 37, 38, 41, 42, 49] aimed to embed images of two modalities into the same feature space to tackle this task, which preliminarily solved the modality gap.

Graph-based methods have achieved excellent performance on standard person ReID [2, 14, 23, 24]. Generally, they predict the pair-wise similarity as relationships between different samples and then utilize those relationships to propagate messages across samples. This kind of methods can bring a large range of performance gain because the features of one sample not only have the discriminative information of this sample itself but also carry information from other relative samples. So several methods [13, 39] attempt to employ graph-based modules to establish relationships and enhance features in the VI-ReID task. But we argue that different from the graph module on the standard Re-ID, the graph methods on VI-ReID suffer bad generalization.

We delve into the graph models pipeline in VI-ReID and summarize two main reasons for the bad generalization: *Train-test modality balance gap* and

*Sub-optimal topology structure.* The train-test modality balance gap is a property of VI-ReID task, which means that the number of two modalities data are balanced in the training stage but extremely unbalanced in inference, like Fig. 1 (a) shows.More details about this property will be further introduced in Section 3. The sub-optimal structure is another problem but usually ignored by previous methods. It is caused by the end-to-end learning manner of the graph model. We summarize that the existing joint learning process of both the backbone and the graph module would weaken the learning of the graph topology structure, like Fig. 1 (b) shows. They both lead to low generalization of the graph structure predicted by the graph module in inference.

To tackle the aforementioned problems respectively, we propose a Counterfactual Intervention Feature Transfer (CIFT) including one new graph module called Homogeneous and Heterogeneous Feature Transfer ($H^2FT$) with one additional learning methods Counterfactual Relation Intervention (CRI). The $H^2FT$ aims to reduce the train-test modality balance gap in two ways, training algorithm and model designing. We reorganize the balanced training data to simulate unbalanced modality distributed scenarios and let the $H^2FT$ trained on that environment, which guides the model to adapt to the situation with unbalanced modality distribution. Also, we find that it is hard for the standard graph module to train efficiently on that unbalanced data because the standard message-passing process cannot adapt to the extremely unbalanced modality information. So, we carefully construct the module of the $H^2FT$ which includes two different types of graph modules, to reduce the useless information introduced by the standard graph module and treat the message passing in unbalanced data better, as Fig. 1 (a) shows. Except that, the CRI tackles the sub-optimal graph topology problem by highlighting the role of graph structure (predicted affinity) in the total end-to-end training. We utilize the tools of causal inference to implement that motivation. We first represent our graph module in the Structural Causal Model [19, 20] in Fig. 1 (b) and modify the training targets of the graph module from only maximizing the probability likelihood to maximizing the combination of both probability likelihood and the total indirect effect (TIE). The former term guides the whole model to classify the identity of each person image. And the latter one is essentially equal to maximize the difference between the original output and a counterfactual output contributed by the affinity changes only (Fig. 1 (b) green background), which can make the model perceive the function of the graph affinity.

The main contributions of our work are summarized as follows:

• We delve into the existing VI-ReID graph model and find two main reasons for their low generalization: train-test modality balance gap and sub-optimal structure. And we design a novel and effective Counterfactual Intervention Feature Transfer (CIFT) to tackle these problems and achieve the new state of the art.

• We introduce a Homogeneous and Heterogeneous Feature Transfer ($H^2FT$) module including two independent types of well-designed graph module and an unbalanced scenario simulation, which is more suitable for tackling the sample interaction in the scenario with unbalanced modality distribution.

• We propose a novel Counterfactual Relation Intervention (CRI) algorithm to tackle the sub-optimal topology structure problem. It utilizes the counterfactual intervention and causal effect tools to highlight the role of the topology in the feature transfer module, which can train the total module more generalized.

## 2 Related Work

**Visible-Infrared Person Re-ID.** Traditional single-modality person Re-ID [15, 26, 43] is limited by the poor illumination conditions at night, so the VI-ReID has received extensive attention in recent years. Many VI-ReID approaches have been proposed to overcome the modality discrepancy produced by different cameras. Wu *et al*. [36] proposed a deep zero-padding network and contribute the first large-scale multiple modality Re-ID dataset named SYSU-MM01.

Many works [7, 10, 11, 37, 38, 41, 42, 49] designed loss functions from the perspective of metric learning to better embed different modalities into the same feature space. Zhu *et al*. [49] proposed the hetero-center loss to reduce the intra-class cross-modality variations. Liu *et al*. [11] proposed the hetero-center triplet loss to relax the strict constraint of traditional triplet loss.

Some methods [4, 31, 32, 34] are based on the generative adversarial network (GAN) [6]. cmGAN [4] adopted generative adversarial training to better distinguish images of different modalities at the feature level. $D^2$RL [34] applied dual-level discrepancy reduction learning based on a bi-directional cycle GAN. Recently, Wu *et al*. [37] introduced a modality alleviation module and a pattern alignment module to discover cross-modality nuances. Hao *et al*. [7] confused two modalities, ensuring that the optimization is explicitly concentrated on the modality-irrelevant perspective. All these methods treat the VI-ReID as an image embedding task and learn to extract features directly from the single image.

**Graph-based Person Re-ID.** In the single-modality person Re-ID task, except for the image embedding method, some works [2, 14, 24, 23] pay attention to the relationship between sample pairs. These methods introduced more supervised information of graph relationships into the training stage, while the inference stage also benefits from pair-wise similarity. Besides, some re-ranking [47] and graph neural networks (GNN) [44] methods only treated relational modeling as post-processing to more flexibly adapt to various backbone networks.

In VI-ReID, the large cross-modality discrepancy makes the optimization of the relationship more difficult. Recently, some approaches have explored cross-modality pair-wise relation learning with graph networks. Ye *et al*. [39] introduced cross-modality graph-structured attention to enhance robustness against noisy samples. Lu *et al*. [13] proposed the cross-modality shared-specific feature transfer algorithm that utilizes the graph convolution operator to propagate features over a graph to supplement the information of another modality. These methods utilized the graph network or transformer module to propagate message cross samples to extract stronger features. But they are all suffering from the train-test modality balance gap and sub-optimal topology problems, which limits their applications. In this paper, we propose a novel graph method CIFT

to tackle these two problems by both model design and learning algorithm, achieving satisfying generalization on VI-ReID.

**Causal Inference in Computer Vision.** The causal inference has recently aroused widespread interest, especially in the combination with computer vision [1, 12, 21, 33] to endow models with the ability to pursue the causal effect. Some works [3, 17, 22, 27, 28, 45] utilized counterfactual to solve problems in various fields of computer vision. Tang *et al.* [27, 28] used counterfactual inference in scene graph generation and long-tailed classification to remove bias from training data with long-tailed distributions. Rao *et al.* [22] used counterfactual training in fine-grained image recognition to tackle the bias of the spatial attention caused by the dataset. Niu *et al.* [17] reduce the language bias in visual question answering by subtracting the direct language effect from the total causal effect.

Different from them, we focus on highlighting the affinity of the feature transfer module to address the sub-optimal topology structure due to the graph-based Re-ID model itself, rather than reducing the impact caused by biased data.

## 3 Delving into Graph-based Visible-Infrared ReID

In this section, we investigate the influence of graph-based modules in the VI-ReID task. Specifically, we first give a brief review of graph-based VI-ReID methods. Then, we investigate why they suffered by bad generalization. Here, we take cm-SSFT [13] and DDAG [39] as examples for analysis.

### 3.1 Review of Graph-based VI-ReID Models

The definition of VI-ReID is essentially a cross-modality retrieval task. So its formula can be written as follow: $\mathcal{R} = \mathcal{M}(q, G)$, where $\mathcal{M}$ is the Re-ID model, used to feedback the ranking list $\mathcal{R}$ between the given query sample $q$ and the gallery set $G$ whose modality is different with the query one. To achieve this pipeline, cm-SSFT [13] and DDAG [39] can be summarized as following:

**Step 1: Modality-invariant feature extraction.** Give an image $x_m$ whatever its modality $m$ ($m \in \{rgb, ir\}$), utilizing CNNs or other backbones to extract features $x$ for each sample.

**Step 2: Feature enhancement.** Build affinities $A$ between all samples in $\{q, G\}$ on their given features $X$, where $A_{i,j}$ means the relationship between the $i$-th and the $j$-th images. After that, messages can be passed and transferred across different samples, leading to stronger features. It can be written as

$$F = A \cdot v(X), \tag{1}$$

where $v$ is a linear learnable function and $F$ stores the output features. This process is essentially equal to constructing a graph whose nodes are person features and edges are affinities and then propagating information based on that graph.

**Step 3: Computing results.** After getting enhanced features, different kinds of outputs, e. g. person identities or ranking lists, can be derived.

**Step 4: Feature learning.** In the training stage, feature learning algorithms are added on both the backbone features $X$ and graph features $F$. The classification output $Y$ is derived by $F$ through a classification layer and a cross-entropy loss is used to train it, which makes features carrying identity information.

The most priority of these graph-based modules is passing messages across samples, which mines the potential relationships between different person images. So, they can benefit both training [13, 39] and inference [39]. But we find that they are all suffering from bad generalization in VI-ReID.

### 3.2 Analysis of Bad Generalization of Graph-based VI-ReID

We summarize that the bad generalization of graph-based VI-ReID is caused by two following problems:

**Train-Test modality balance gap.** The train-test modality balance gap is caused by the difference of modality information ratio in the training and test stage. Specifically, in the training stage, both cm-SSFT [13] and DDAG [39] pass messages and transfer features on the batch data which includes an equal number of visible and infrared images. So the ratio of two modality information provided in training is $1:1$. But in inference, the available data is $\{q, G\}$ consisting of one query sample $q$ and a gallery set $G$. Here, the modality information ratio between two modalities is $1:N_G$, where $N_G$ is the size of the gallery set. It is clear that the modality information ratio of training and testing is quite different. This is the property of VI-ReID because VI-ReID utilizes the single query evaluation setting which means there is only one query sample available in the inference scenario. It is hard for the model trained on the balanced training data to generalize on the unbalanced inference scenario. The cm-SSFT [13] also provides a series of experiments that demonstrate the unbalanced inference scenario can actually harm the generalization, which brings about 13.9% Rank-1 and 9.1% mAP drops corresponding to a balanced inference one.

**Sub-optimal topology structure.** The affinities $A$ computed by cm-SSFT [13] and DDAG [39] can indicate the relationships between different samples. So the $A$ can be interpreted as a kind of graph topology structure on the given data. But we argue that the structure learned by the existing graph VI-ReID modules are all sub-optimal because of the end-to-end joint learning.

Both cm-SSFT [13] and DDAG [39] add supervisions on backbone features and graph features simultaneously without any constraint on the affinities $A$, which hurts the generalization. It is common for the graph modules, like transformer [29] or Graph Attention [30], to train the $A$ in an end-to-end joint manner. But the situation is different here, the supervision on the backbone makes the backbone features $X$ discriminative in the training set. At this time, an $A$ with standard quality can make the final output belong to the feature learning constraints[1], so the structure $A$ cannot get much useful guidance. Without further supervision of $A$, it is hard for the graph module to capture the complex relationships between different samples.

[1] Further proves could be seen in the supplementary.

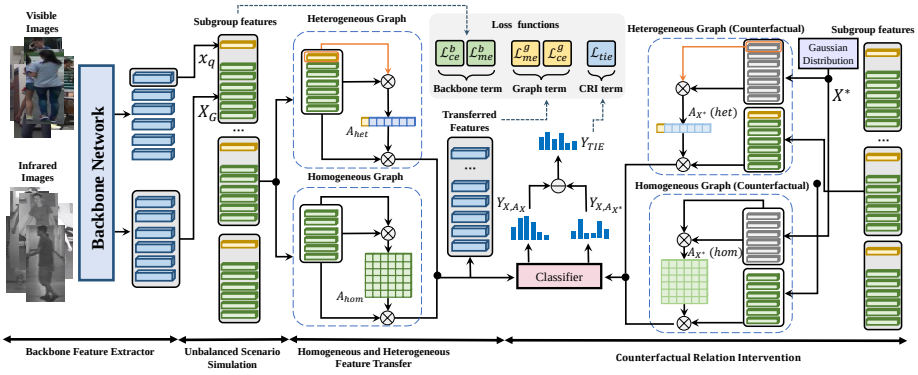


**Fig. 2.** The Framework of the proposed Counterfactual Intervention Feature Transfer (CIFT). The Homogeneous and Heterogeneous Feature Transfer ($H^2FT$) module receives features from the backbone feature extractor and simulates unbalanced modality distributed scenarios. It builds homogeneous and heterogeneous message passing for better relationship feature learning under the unbalanced modality distribution. Then, the Counterfactual Relation Intervention (CRI) module calculates total indirect effects to highlight the role of the graph topology and leads to stronger results.

The above analyses reveal that the key to increasing the generalization of graph-based VI-ReID is reducing the modality balance gap between train-test and introducing additional constraints on $A$ in the end-to-end joint learning. Along this direction, we proposed a Counterfactual Intervention Feature Transfer module and show how this model is used to tackle these problems.

## 4 Counterfactual Intervention Feature Transfer

The overview of our proposed Counterfactual Intervention Feature Transfer (CIFT) is shown in Fig. 2. Visible images and infrared images are first fed into a pedestrian feature extractor to extract the instance-level features. Then these features are sent to the graph module Homogeneous and Heterogeneous Feature Transfer ($H^2FT$) to extract transferred features (§ 4.2). Meanwhile, the Counterfactual Relation Intervention (CRI) is introduced to improve graph structure learning (§ 4.3). The optimization and loss function details are shown in (§ 4.4).

### 4.1 Pedestrian representation backbone

As shown in Fig. 2, our pedestrian representation backbone network is a weight sharing network for both modality data, which embeds the person images from different modalities to a same feature space. To make sure the representation ability of those features, a loss function $\mathcal{L}_{ce}^{b}$ including a classification loss and a cross-modality metric learning loss $\mathcal{L}_{me}^{b}$ is utilized to optimize the networks. Based on that, each person's image can have an initial representation.

### 4.2 Homogeneous and Heterogeneous Feature Transfer

For a given batch data including $N$ Visible and $N$ infrared images, we aim to simulate the unbalanced modality distributed scenario to train our model. The batch data is split into a series of groups and each group consists of a single image from one modality (seen as the query $q$) and $N$ other modality images (seen as the gallery set $G$). Specifically, each group can have one visible image with $N$ infrared images or one infrared image with $N$ visible images. Every group simulates the scenario with $1:N$ modality ratio which is similar in the inference setting (one query vs more galleries). So, our H$^2$FT trained under that can adapt on the single query inference without much influence on generalization.

But, under that setting, the modality information is quite imbalanced. The sample seen as the query can only interact with $N$ images from other modality. And for the data seen as gallery, inter-modality interaction is trivial because it just introduces the information provided by the fixed query one, which is redundant even noisy. To avoid this problem, we provide a heterogeneous and a homogeneous graph module for the query data and gallery data separately like Fig. 2 shows. Its equation can be written as:

$$f_q = A_{het} \cdot [v(x_q), v(X_G)],\ F_G = A_{hom} \cdot v(X_G), \tag{2}$$

where $x_q$ and $X_G$ mean the query feature vector and the gallery features matrix respectively. $[\bullet, \bullet]$ means concatenation in the column dimension. The function $v(\cdot)$, a BNNeck [15] with learnable weights, is used to enhanced the features. $f_q$ and $F_G$ mean the transferred features of the query and the gallery set. $A$ means affinity matrix indicating the relationships between its corresponding samples. To achieve $A$, we first compute the similarity matrix based on the input features:

$$\mathcal{S}_{ty}^{i,j} = \exp \frac{cos(v(x_i), v(x_j))}{\tau_{ty}}, \quad ty \in \{hom, het\}. \tag{3}$$

$cos(\cdot, \cdot)$ is the cosine distance function to measure the similarity between samples. $\tau$ is the temperature parameter to adjust the smoothness of the total similarity distribution. We use different $\tau$ for the heterogeneous and homogeneous process because the intra-modality and inter-modality similarities are quite different. To filter out the noisy relationships, we use near neighbor chosen function $\mathcal{T}(\bullet, k)$ [13] to keep the top-$k$ values in each row of similarity matrix: $\mathbf{S}' = \mathcal{T}(\mathbf{S}, k)$. Finally, our affinity matrices are computed as follows:

$$A_{hom} = \mathbf{D}_{hom}^{-1} \cdot \mathbf{S}'_{hom},\ A_{het} = \mathbf{D}_{het}^{-1} \cdot \mathbf{S}'_{het}. \tag{4}$$

$\mathbf{D}^{-1}$ is the Laplacian matrix of the $\mathbf{S}'$, used to normalize the total affinities.

Finally, after achieving the final transferred features $F$, a classification layer derives the logits $Y$ and the cross-entropy loss $\mathcal{L}_{ce}^{g}$ is utilized to train $Y$ which is equal to maximizing likelihood to keep the features carrying richer identity information. Then metric learning term $\mathcal{L}_{me}^{g}$ is added to the output to make features carry more discriminative information.

Along that pipeline, the unbalanced scenario simulation makes the model adapt to the single query inference scenario. And the two kinds of message passing processes, heterogeneous and homogeneous, guide the query sample to interact with potential galleries while the gallery data only propagate messages across themselves, which preserves the nontrivial information interaction process in each group data, which is more suitable for tackling this unbalanced situation.

### 4.3 Counterfactual Relation Intervention

To add additional supervision of the affinities $A$ and keep the whole end-to-end training pipeline, we present to highlight the role of the graph topology structure in the total learning process. For this goal, we bring the tools of causal inference here. We first represent our H$^2$FT into a Structural Causal Model (SCM) [19, 20], like Fig. 1 (b) shows. $X \rightarrow A$ means the affinity computation and $X \rightarrow Y \leftarrow A$ is the message passing process (including the output computation).

It is obvious the process that deriving the output $Y$ from input $X$ can be seen as two types of effects: One is the direct effect $X \rightarrow Y$ and the other is an indirect one $X \rightarrow A \rightarrow Y$. The classification loss for our H$^2$FT, equal to maximizing the likelihood, would affect the two effects in an end-to-end manner so the $A$ in the indirect effect path cannot be enhanced sufficiently.

To highlight the $A$ in the whole training process, we utilize the Total Indirect Effect (TIE) here. We first give its equation:

$$Y_{TIE} = Y_{X,A_X} - \mathbb{E}_{X^*}[Y_{X,A_{X^*}}]. \tag{5}$$

$Y_{X,A_X}$ is the original output of our graph module, which means feeding forward the different sample features $X$ and computing their outputs. Note that the affinity matrix here is denoted as $A_X$ which means that affinities are computed based on the input features $X$. $Y_{X,A_{X^*}}$ means computing the results by replacing original affinity $A_X$ to a intervened one $A_{X^*}$, where $X^*$ is the intervened inputs given manually. It is obvious that $Y_{X,A_{X^*}}$ cannot occur in the real world because features $X$ and affinities $A_{X^*}$ come from different inputs $X$ and $X^*$, which is called counterfactual intervention. So modification from $Y_{X,A_X}$ to $Y_{X,A_{X^*}}$ is equal to keep all potential variables fixed but only change the affinity $A$, which can show the pure effect introduced by $A$. We compute the expectation of that effect to get the more stable one. Intervened input features $X^*$ utilized to compute $A_{X^*}$ are sampled by a Gaussian distribution:

$$X^* = X_\sigma \cdot Z + X_\mu, \tag{6}$$

where $Z$ is the standard random vector whose dimension is same with features $X$. mean $X_\mu$ and stand deviation $X_\sigma$ are learned by the re-parameterization trick [9] in an end-to-end way.

A cross-entropy loss is added to the TIE: $\mathcal{L}_{tie} = \mathcal{L}_{ce}(Y_{TIE})$. Minimizing that cross-entropy loss is equal to maximizing the $Y_{TIE}$ on the prediction of the correct class, which guides the model to increase the gap between the original output

and the counterfactual one. It is clear that the counterfactual classification results should be worse than the original one because the intervened affinities $A_{X^*}$ commonly do not match with inputs $X$. So an intuitive understanding about maximizing TIE is constraining the model to increase the difference between the outputs derived from good $A_X$ and bad $A_{X^*}$. Since other variables have been fixed, the model has to change the $A_X$ to increase the original results $Y_{X,A_X}$ for enhancing the gaps, leading to better training of affinity.

### 4.4 Optimization

The whole model is trained end-to-end and the total loss $\mathcal{L}_{total}$ of our method is defined as:[2]

$$\mathcal{L}_{total} = \underbrace{\mathcal{L}^b_{ce} + \mathcal{L}^b_{me}}_{\text{backbone term}} + \underbrace{\mathcal{L}^g_{ce} + \mathcal{L}^g_{me}}_{\text{graph term}} + \underbrace{\mathcal{L}_{tie}}_{\text{CRI term}} . \tag{7}$$

## 5 Experiments

### 5.1 Datasets and Evaluation Protocol

In this section, we conduct comprehensive experiments to evaluate our method on two public datasets, SYSU-MM01 [36] and RegDB [16].

**SYSU-MM01** is the first large-scale benchmark dataset for Visible-Infrared ReID. It is collected by four visible and two infrared cameras, in both indoor and outdoor environments. The training set contains 395 identities with 22,258 visible and 11,909 infrared images while the test set contains 96 identities. Concretely, the query set contains 3,803 infrared images and the gallery set contains 301/3010 (single-shot/multi-shot) randomly selected visible images.

**RegDB** is collected by a dual-camera system(a pair of aligned visible and thermal cameras). It contains 412 people, and each person has 10 visible and 10 far-infrared images. The dataset is divided into training and test splits randomly, the images of 206 identities for training and the rest 206 identities for testing.

**Evaluation Protocol.** All the experiments follow the standard evaluation protocol in existing Visible-Infrared cross-modality ReID benchmarks. For SYSU-MM01, the original evaluation protocol [36] provides all-search and indoor-search modes for testing. Both search modes have two retrieval settings, single-shot and multi-shot. For RegDB, we follow the widely used evaluation protocol in [16] which contains two modes for testing, Visible to Infrared test mode and Infrared to Visible test mode. We evaluate our model on the 10 trials with different training/test splits to achieve stable performance. For both datasets, the cumulative matching characteristics (CMC) and mean average precision (mAP) are adopted as evaluation metrics.

[2] The details about cross-modality metric learning loss $\mathcal{L}_{me}$ and $\mathcal{L}_{me}$ can be found in the supplementary.

**Table 1.** Comparison of rank-1 accuracy (%) and mAP accuracy(%) with the state-of-the-art methods on SYSU-MM01 and RegDB. (CIFT† means we use backbone features for inference rather than the transferred graph features.)

| Method | SYSU-MM01 [36] | | | | | | | | RegDB [16] | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | All-search | | | | Indoor-search | | | | Visible to Infrared | | Infrared to Visible | |
| | Single-shot | | Multi-shot | | Single-shot | | Multi-shot | | | | | |
| | rank-1 | mAP | rank-1 | mAP | rank-1 | mAP | rank-1 | mAP | rank-1 | mAP | rank-1 | mAP |
| Zero-Pad [36] | 14.80 | 15.95 | 19.13 | 10.89 | 20.58 | 26.92 | 24.43 | 18.64 | - | - | - | - |
| cmGAN [4] | 26.97 | 27.80 | 31.49 | 22.27 | 31.63 | 42.19 | 37.00 | 32.76 | - | - | - | - |
| $D^2$RL [34] | 28.9 | 29.2 | - | - | - | - | - | - | 43.4 | 44.1 | - | - |
| JSIA-ReID [31] | 38.1 | 36.9 | 45.1 | 29.5 | 43.8 | 52.9 | 52.7 | 42.7 | 48.5 | 49.3 | 48.1 | 48.9 |
| AlignGAN [32] | 42.4 | 40.7 | 51.5 | 33.9 | 45.9 | 54.3 | 57.1 | 45.3 | 57.9 | 53.6 | 56.3 | 53.4 |
| AGW [40] | 47.5 | 47.65 | - | - | 54.17 | 62.97 | - | - | 70.05 | 66.37 | - | - |
| cm-SSFT [13] | 61.6 | 63.2 | 63.4 | 62.0 | 70.5 | 72.6 | 73.0 | 72.4 | 72.3 | 72.0 | 71.0 | 71.7 |
| cm-SSFT(sq) | 47.7 | 54.1 | - | - | 57.4 | 59.1 | - | - | 65.4 | 65.6 | 63.8 | 64.2 |
| DDAG [39] | 54.75 | 53.02 | - | - | 61.02 | 67.98 | - | - | 69.34 | 63.46 | 68.06 | 61.80 |
| HC [49] | 56.96 | 54.95 | 62.09 | 48.02 | 59.74 | 64.91 | 69.76 | 57.81 | - | - | - | - |
| CIMA [46] | 57.2 | 59.3 | 60.7 | 52.6 | 66.6 | 74.7 | 73.8 | 68.3 | 78.8 | 69.4 | 77.9 | 69.4 |
| HCT [11] | 61.68 | 57.51 | - | - | 63.41 | 68.17 | - | - | 91.05 | 83.28 | 89.30 | 81.46 |
| MCLNet [7] | 65.4 | 61.98 | - | - | 72.56 | 76.58 | - | - | 80.31 | 73.07 | 75.93 | 69.49 |
| SMCL [35] | 67.39 | 61.78 | 72.15 | 54.93 | 68.84 | 75.56 | 79.57 | 66.57 | 83.93 | 79.83 | 83.05 | 78.57 |
| MPANet [37] | 70.58 | 68.24 | 75.58 | 62.91 | 76.74 | 80.95 | 84.22 | 75.11 | 83.7 | 80.9 | 82.8 | 80.7 |
| CIFT†(Ours) | 71.77 | 67.64 | 78.00 | 62.46 | 78.65 | 82.11 | 86.97 | 77.03 | **92.17** | 86.96 | 90.12 | 84.81 |
| CIFT(Ours) | **74.08** | **74.79** | **79.74** | **75.56** | **81.82** | **85.61** | **88.32** | **86.42** | 91.96 | **92.00** | **90.30** | **90.78** |

### 5.2 Implementation Details

We implement our approach with PyTorch [18] on one NVIDIA Titan Xp GPU. Following the previous ReID methods [15, 40], we use ResNet-50 [8] pre-trained on ImageNet as our backbone network. We change the stride of the last convolutional layer in the backbone to 1 and employ the Batch Normalization Neck [15] as the embedding layer. Each person image is resized to commonly used $288 \times 144$ resolution. We also adopt the random cropping, random horizontal flipping and random erasing [48] for data augmentation. The $k$ in the near neighbor chosen function is set to 4. $\tau_{hom}$ and $\tau_{het}$ in Eq. 3 are set to 0.4 and 0.2 for the heterogeneous and homogeneous process. The whole model is trained for 120 epochs with the SGD optimizer. The learning rate gradually rises up by the warm-up scheme and decays by a factor of 10 at the 60th and 100th epochs. The batch size is set to 64, containing 32 visible and 32 infrared images from 8 identities. And each identity consists of 4 visible and 4 infrared images.

### 5.3 Comparison with State-of-the-art Methods

In this part, we compare our proposed method CIFT with state-of-the-art (SOTA) visible-infrared person Re-ID approaches, including Zero-Pad [36], cmGAN [4], $D^2$RL [34], JSIA-ReID [31], AlignGAN [32], AGW [40], cm-SSFT [13], DDAG [39], HC [49], CIMA [46], HCT [11], MCLNet [7], SMCL [35] and MPANet [37].

**Comparison and Analysis.** The experimental results are shown in Table 1 and the proposed method outperforms the existing SOTAs on both datasets. In SYSU-MM01 dataset, our CIFT achieves 74.08% rank-1 accuracy and 74.79% mAP accuracy, which surpasses MPANet [37] by 3.50% on rank-1 accuracy and 6.55% on mAP accuracy in the most challenging single-shot all search mode. Even our CIFT$^{\dagger}$ (only uses graph module in training and utilizes the backbone features in inference) outperforms MPANet by 1.19% on rank-1 accuracy. In another popular public RegDB dataset, whether in 'infrared to visible' mode or 'visible to infrared' mode, our CIFT$^{\dagger}$ still achieves the highest scores. The average performances in the two modes are 91.15% rank-1 accuracy and 85.89% mAP accuracy, which surpasses MPANet by a large gain of about 7.90% on rank-1 accuracy and 5.09% on mAP accuracy. This is because the learning of transferred features introduces additional supervision to the model, so that the features of the backbone network are also enhanced.

For the multi-shot setting, the mAP accuracy of all other methods will drop significantly compared with the single-shot evaluation because the model is required to retrieve more positive targets in the multi-shot setting. So it is more challenging for the model to find out all potential targets. But our CIFT is not suffering that bad phenomenon even can achieve better results, which shows that our method is qualitatively different from other methods. When the scale gallery size is larger, our model can extract richer and more discriminate relation features so that the model is more robust to gallery size even benefited by the larger scale one. Compared with cm-SSFT [13] which also obtain relationship from the gallery set, our CIFT improves the mAP accuracy by 0.77% as the gallery size increases (multi-shot versus single-shot), while the cm-SSFT reduces the mAP accuracy by 1.2%. This is because we utilize CRI to highlight the role of topology structure in the whole training process. The affinities are much more accurate so that the final representation is much stronger.

### 5.4 Comparison with Multi-gallery Matching Methods

To demonstrate the superiority of our CIFT to other graph and post-process VI-ReID methods, we compare it with other multi-gallery matching methods, including cm-SSFT [13], k-reciprocal rerank [47], and GNN rerank [44]. We evaluate these methods with our CIFT on different backbone networks including AGW [40], HC [49], HCT [11], and our backbone network, to show the generality of our method under different level baselines. For a fair comparison, we also search the best hyper-parameters for these multi-gallery matching methods, so that they can fit the backbones well. Please note that, to adapt these backbone features format, cm-SSFT is set as only using the shared feature transfer.

Our results are shown in the 5th line in Table 2 and 1st∼3rd lines represent different widely used post-processing ways, combining different multi-gallery matching methods with the trained backbone features directly in inference. Comparing with the strongest post-process GNN rerank [44], we achieve averaged 5.12% rank-1 and 4.03% mAP gains on all given backbones. Further, we also compare the single query cm-SSFT (the 4th row) who does additional feature

**Table 2.** Compared with other methods that use gallery set information in the inference stage on same backbone networks, *i.e.* AGW [40], HC [49], HCT [11] and our backbone. We report the rank-1 accuracy (%) and the mAP accuracy (%) on the SYSU-MM01 single-shot all search mode. 'train' and 'test' in 'strategy' means the training methods and test methods separately

| Row | backbone / Strategy | AGW [40] | | HC [49] | | HCT [11] | | Our backbone | |
|---|---|---|---|---|---|---|---|---|---|
| | | rank-1 | mAP | rank-1 | mAP | rank-1 | mAP | rank-1 | mAP |
| 1 | backbone | 47.22 | 47.78 | 54.52 | 54.06 | 61.05 | 56.97 | 70.49 | 66.58 |
| 2 | k-reciprocal [47] | 47.63 | 51.81 | 53.91 | 60.03 | 62.33 | 62.05 | 71.47 | 72.49 |
| 3 | GNN rerank [44] | 46.96 | 52.01 | 55.05 | 59.79 | 60.52 | 62.72 | 70.40 | 73.21 |
| 4 | cm-SSFT [13] | 48.65 | 51.76 | 56.17 | 61.36 | 62.27 | 61.85 | 69.92 | 71.77 |
| 5 | **CIFT** | **52.12** | **56.92** | **61.03** | **64.05** | **66.18** | **68.09** | **74.08** | **74.79** |

**Table 3.** Ablation study on SYSU-MM01. The important modules of the proposed CIFT, *i.e.* $H^2$FT and CRI are analyzed under different settings.

| Row | GFT | UBS | $H^2$G | CRI | SYSU-MM01 | |
|---|---|---|---|---|---|---|
| | | | | | rank-1 | mAP |
| 1 | - | - | - | - | 70.49 | 66.58 |
| 2 | ✓ | - | - | - | 72.01 | 72.12 |
| 3 | ✓ | ✓ | - | - | 72.90 | 71.97 |
| 4 | ✓ | ✓ | ✓ | - | 72.29 | 73.79 |
| 5 | ✓ | ✓ | ✓ | ✓ | 74.08 | 74.79 |

**Table 4.** Affinity quality statistics on the SYSU-MM01 test set. The value (%) represents the **average error ratio** of the affinity matrix in the entire test set.

| Method | All-search | | Indoor-search | |
|---|---|---|---|---|
| | Single-shot | Multi-shot | Single-shot | Multi-shot |
| w/o CRI | 5.16 | 3.95 | 6.54 | 6.15 |
| w/ CRI | 3.90 | 2.76 | 5.03 | 4.66 |

transfer learning corresponding to the aforementioned post-process methods. It only achieves comparable results with GNN rerank in the 3rd line and does not show much more priority of its graph learning. That is because cm-SSFT trains in the case of balanced modality but transferred to modal unbalanced inference scenario, which hurts its generalization. So, with the ability to tackle the inference under unbalanced modality distribution, it is common for CIFT to suppress the cm-SSFT by a large margin. The results in Table 2 show that our method achieves the best performance on all backbone networks, bringing average improvements of 5.03% on rank-1 accuracy and 9.50% on mAP accuracy. This proves that our method is compatible with various backbone networks and can achieve effective improvement.

### 5.5 Ablation Study

In this section, we conduct ablation studies to prove the effectiveness of each module of the proposed CIFT, *i.e.* $H^2$FT and CRI. All ablation experiments are performed on our baseline backbone in the single-shot all search mode of the large-scale dataset SYSU-MM01. The results are shown in Table 3.

**Effectiveness of $H^2$FT.** In the proposed CIFT, we introduce a graph-based feature transfer module $H^2$FT to tackle the train-test modality balance gap. To evaluate the effectiveness of each detailed part in $H^2$FT, We split the $H^2$FT into three parts: Graph Feature Transfer (GFT), UnBalanced Scenario simulation (UBS) and Homogeneous&Heterogeneous Graph module ($H^2$G). GFT is a simple graph module baseline proposed by ourselves, which is used to show the gain introduced by the message passing in the graph module itself. Its details can be seen in the supplementary and we try ourselves to keep other variables not changing. Its performance in the 2nd row shows that the feature transfer can bring about 5.5% gains in mAP. To train the model suitable for the unbalanced inference scenario, we add the UBS on it. But we find that the performance on mAP has a little drop. We think that is caused by the model design who is not fit the unbalanced data. Now, we add the $H^2$G back in the 3rd line, equal to the complete $H^2$FT, and achieve additional 1.67% mAP gains. proving the effectiveness of our $H^2$FT.

**Effectiveness of CRI.** The proposed CRI algorithm utilizes the counterfactual intervention to highlight the role of the topology to tackle the sub-optimal topology structure problem. As shown in Table 3, CRI algorithm brings improvements of 1.79% rank-1 accuracy and 1.00% mAP accuracy without any computational costs in the inference stage. In addition, we also do another quantitative analysis of the CRI to demonstrate its contribution. Specifically, we first compute the ideal affinity matrix by the ground-truth label, which uses 1 to indicate the positive pair and 0 as the negative one. And then for the affinity matrix computed by our model, we define the top-4 results in each row as positive. After that, we compute the averaged error rate between the predicted affinities with the ground truths in the entire test set. The results are shown in Table 4. In the single-shot all-search mode, the model without CRI gets for 5.16% error ratio while the model with CRI achieves 3.90%. In other test modes, introducing CRI also significantly reduces the error ratios. This reflects that our CRI has learned a better structure which is more close to the ground-truth one.

## 6 Conclusion

We propose a Homogeneous and Heterogeneous Feature Transfer ($H^2$FT) module with a Counterfactual Relation Intervention (CRI) learning method to tackle the Visible-Infrared Person Re-identification. The $H^2$FT consists of two types of graph modules that can handle the train-test modality balance gap that the previous graph-based model suffered. And CRI introduces the causal inference tool to tackle the sub-optimal topology structure problem and makes our method more generalized.

## References


1. Chalupka, K., Perona, P., Eberhardt, F.: Visual causal feature learning. arXiv preprint arXiv:1412.2309 (2014)
2. Chen, D., Xu, D., Li, H., Sebe, N., Wang, X.: Group consistent similarity learning via deep crf for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8649–8658 (2018)
3. Chen, L., Yan, X., Xiao, J., Zhang, H., Pu, S., Zhuang, Y.: Counterfactual samples synthesizing for robust visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10800–10809 (2020)
4. Dai, P., Ji, R., Wang, H., Wu, Q., Huang, Y.: Cross-modality person re-identification with generative adversarial training. In: IJCAI. vol. 1, p. 2 (2018)
5. Gong, S., Cristani, M., Loy, C.C., Hospedales, T.M.: The re-identification challenge. In: Person re-identification, pp. 1–20. Springer (2014)
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
7. Hao, X., Zhao, S., Ye, M., Shen, J.: Cross-modality person re-identification via modality confusion and center aggregation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16403–16412 (2021)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
10. Ling, Y., Luo, Z., Lin, Y., Li, S.: A multi-constraint similarity learning with adaptive weighting for visible-thermal person re-identification. In: IJCAI. pp. 845–851 (2021)
11. Liu, H., Tan, X., Zhou, X.: Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification. IEEE Transactions on Multimedia (2020)
12. Lopez-Paz, D., Nishihara, R., Chintala, S., Scholkopf, B., Bottou, L.: Discovering causal signals in images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6979–6987 (2017)
13. Lu, Y., Wu, Y., Liu, B., Zhang, T., Li, B., Chu, Q., Yu, N.: Cross-modality person re-identification with shared-specific feature transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13379–13389 (2020)
14. Luo, C., Chen, Y., Wang, N., Zhang, Z.: Spectral feature transformation for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4976–4985 (2019)
15. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
16. Nguyen, D.T., Hong, H.G., Kim, K.W., Park, K.R.: Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors **17**(3), 605 (2017)
17. Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X.S., Wen, J.R.: Counterfactual vqa: A cause-effect look at language bias. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12700–12710 (2021)

18. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
19. Pearl, J., Glymour, M.a., Jewell, N.P.: Causal inference in statistics: A primer. John Wiley & Sons (2016)
20. Pearl, J., Mackenzie, D.: The book of why: the new science of cause and effect. Basic books (2018)
21. Qi, J., Niu, Y., Huang, J., Zhang, H.: Two causal principles for improving visual dialog. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10860–10869 (2020)
22. Rao, Y., Chen, G., Lu, J., Zhou, J.: Counterfactual attention learning for fine-grained visual categorization and re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1025–1034 (2021)
23. Shen, Y., Li, H., Xiao, T., Yi, S., Chen, D., Wang, X.: Deep group-shuffling random walk for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2265–2274 (2018)
24. Shen, Y., Li, H., Yi, S., Chen, D., Wang, X.: Person re-identification with deep similarity-guided graph neural network. In: Proceedings of the European conference on computer vision (ECCV). pp. 486–504 (2018)
25. Sun, Y., Cheng, C., Zhang, Y., Zhang, C., Zheng, L., Wang, Z., Wei, Y.: Circle loss: A unified perspective of pair similarity optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6398–6407 (2020)
26. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European conference on computer vision (ECCV). pp. 480–496 (2018)
27. Tang, K., Huang, J., Zhang, H.: Long-tailed classification by keeping the good and removing the bad momentum causal effect. arXiv preprint arXiv:2009.12991 (2020)
28. Tang, K., Niu, Y., Huang, J., Shi, J., Zhang, H.: Unbiased scene graph generation from biased training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3716–3725 (2020)
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
30. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)
31. Wang, G.A., Zhang, T., Yang, Y., Cheng, J., Chang, J., Liang, X., Hou, Z.G.: Cross-modality paired-images generation for rgb-infrared person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12144–12151 (2020)
32. Wang, G., Zhang, T., Cheng, J., Liu, S., Yang, Y., Hou, Z.: Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3623–3632 (2019)
33. Wang, T., Huang, J., Zhang, H., Sun, Q.: Visual commonsense r-cnn. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10760–10770 (2020)
34. Wang, Z., Wang, Z., Zheng, Y., Chuang, Y.Y., Satoh, S.: Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 618–626 (2019)

35. Wei, Z., Yang, X., Wang, N., Gao, X.: Syncretic modality collaborative learning for visible infrared person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 225–234 (2021)
36. Wu, A., Zheng, W.S., Yu, H.X., Gong, S., Lai, J.: Rgb-infrared cross-modality person re-identification. In: Proceedings of the IEEE international conference on computer vision. pp. 5380–5389 (2017)
37. Wu, Q., Dai, P., Chen, J., Lin, C.W., Wu, Y., Huang, F., Zhong, B., Ji, R.: Discover cross-modality nuances for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4330–4339 (2021)
38. Ye, M., Lan, X., Li, J., Yuen, P.: Hierarchical discriminative learning for visible thermal person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
39. Ye, M., Shen, J., J. Crandall, D., Shao, L., Luo, J.: Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16. pp. 229–247. Springer (2020)
40. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.: Deep learning for person re-identification: A survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
41. Ye, M., Wang, Z., Lan, X., Yuen, P.C.: Visible thermal person re-identification via dual-constrained top-ranking. In: IJCAI. vol. 1, p. 2 (2018)
42. Zhang, L., Du, G., Liu, F., Tu, H., Shu, X.: Global-local multiple granularity learning for cross-modality visible-infrared person reidentification. IEEE Transactions on Neural Networks and Learning Systems (2021)
43. Zhang, X., Luo, H., Fan, X., Xiang, W., Sun, Y., Xiao, Q., Jiang, W., Zhang, C., Sun, J.: Alignedreid: Surpassing human-level performance in person re-identification. arXiv preprint arXiv:1711.08184 (2017)
44. Zhang, X., Jiang, M., Zheng, Z., Tan, X., Ding, E., Yang, Y.: Understanding image retrieval re-ranking: a graph neural network perspective. arXiv preprint arXiv:2012.07620 (2020)
45. Zhang, Z., Zhao, Z., Lin, Z., He, X., et al.: Counterfactual contrastive learning for weakly-supervised vision-language grounding. Advances in Neural Information Processing Systems **33**, 18123–18134 (2020)
46. Zhao, Z., Liu, B., Chu, Q., Lu, Y., Yu, N.: Joint color-irrelevant consistency learning and identity-aware modality adaptation for visible-infrared cross modality person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 3520–3528 (2021)
47. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1318–1327 (2017)
48. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 13001–13008 (2020)
49. Zhu, Y., Yang, Z., Wang, L., Zhao, S., Hu, X., Tao, D.: Hetero-center loss for cross-modality person re-identification. Neurocomputing **386**, 97–109 (2020)