Counterfactual Intervention Feature Transfer for Visible-Infrared Person Re-identification (Supplementary Material)

1 Further proof about sub-optimal topology structure

The motivation of the Counterfactual Relation Intervention (CRI) is that joint learning of graph inputs X and outputs Y leads to bad affinities A. Here, we do some toy example experiments to demonstrate that phenomenon. We aim to investigate the relationships of A with the quality of X and Y. For example, if the quality of X and Y are good (meaning both graph and backbone features trained well), how about the quality of A. We firstly define the metric to evaluate features and affinities. For a good learned representation, features should fit the following scheme:

$$\begin{aligned} GR(X^i) &= \mathcal{D}(X^i, X^k) - \mathcal{D}(X^i, X^j), \quad \forall j \in \mathcal{P}^i, k \in \mathcal{N}^i, \\ GR(Y^i) &= \mathcal{D}(Y^i, Y^k) - \mathcal{D}(Y^i, Y^j), \quad \forall j \in \mathcal{P}^i, k \in \mathcal{N}^i. \end{aligned} \tag{1}$$

j and k mean the indexes of the positive set $\mathcal{P}^i$ and the negative set $\mathcal{N}^i$ for the i-th samples. $\mathcal{D}$ is the distance function. These equations compute the margins between distances of its all positive pairs and their corresponding negative ones. We statistic the averaged margin as the quality metrics:

$$Q_X = \frac{\sum i \in \{GR(X^i)\}}{N}, \quad Q_Y = \frac{\sum i \in \{GR(Y^i)\}}{N}, \tag{2}$$

where N is the number of features, equal for the graph one and the backbone one. And we also propose a metric to evaluate affinities A:

$$Q_A = \frac{\sum_{i \in GA(A^i)} i}{N}, \quad GA(A^i) = \{i, A^{i,j} > A^{i,k}\}, \quad j \in \mathcal{P}^i, k \in \mathcal{N}^i, \tag{3}$$

where $\{i, cod\}$ means the set of i who satisfies the condition cod. That equation means that the good affinities should include larger positive similarities than corresponding negative ones. So this Q_A metric measure the ratio of samples' affinities belonging to that constraint.

After that, we randomly generate a series of triplets $\{X, A, Y\}$ and control their quality carefully. And then, we evaluate the quality of Q_Y and draw these data in Figure 1 and see how do the Q_X and Q_A affect Q_Y. To achieve stable results, we reproduce that statistic 100 times and compute mean results. From Figure 1 (a) (bird of view figure), it obvious that A cannot learn sufficient when the inputs features are trained well. A low-quality affinities A (low Q_A, e.g. about 0.1, the blue dash line in Figure 1 (a)) can also get a good transferred

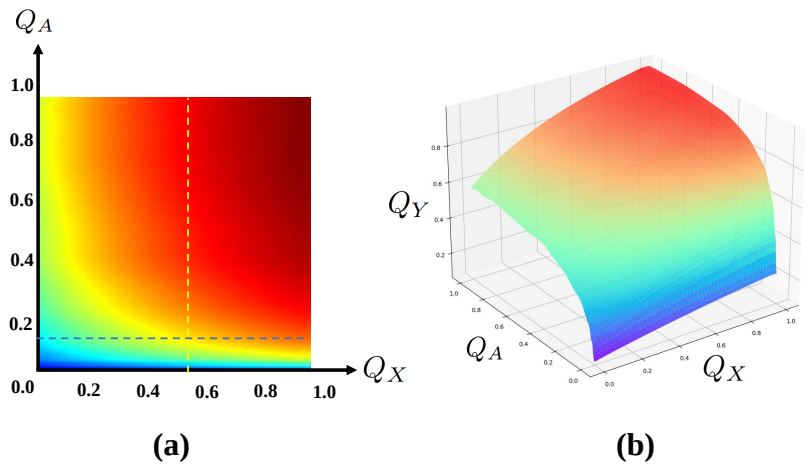

Fig. 1. (a) shows the relationships between Q_A, Q_X and Q_Y. The blue dash line gives an example about fixing Q_A and seeing the changes of Q_Y introduced by Q_X. The yellow dash one aims to analysis the changes from Q_A to Q_Y on the real fixed Q_A computed on our baseline on the SYSU-MM01 benchmark. (b) is the 3D surface version of (a).

features Y (high Q_Y) as long as having a high-quality input representation X (high Q_X). It proves that if the A is not good, the graph output features Y can also have a good representation abilities because of good X. So the training of X can relax the constraint of A. And as we know, in the training set of ReID task, features also trained well. So the range of A could be flexible. To prove that, we compute the Q_A on our baseline network and the result is 0.57. We analysis the Q_A influence to Q_Y on this constant Q_X value (yellow dash line in Figure 1 (a)). And we find that, the range of A is extreme big. As long as the Q_A upper than 0.16, the outputs of Y can achieve satisfied quality, larger than 0.7. It shows that well-trained X and Y can lead to a flexible range of A, bringing sub-optimal topology structure.

2 Feature learning loss function details

For both backbone and graph module features, we add cross-entropy losses to train the features include identity information:

$$\mathcal{L}_{ce}^{b}(Y_b) = \mathbb{E}_i[-\log(Y_b^i)], \quad \mathcal{L}_{ce}^{g}(Y_g) = \mathbb{E}_i[-\log(Y_g^i)], \tag{4}$$

where $Y_\cdot^i$ is the predicted probability for the ground-truth category of the i-th sample. It is computed based on the features $F_\cdot^i$ by the corresponding classification layer:

$$Y_b^i = W_b F_b^i, \quad Y_g^i = W_g F_g^i. \tag{5}$$

where $W_{\cdot}$ means learnable parameters of classification layers. Except that, we design a new metric learning loss called Heterogeneous Center Contrastive (HCC) loss:

$$\mathcal{L}_{hcc}(I, C, \mathcal{D}) = \mathbb{E}_i\{\mathcal{D}(I^i, C^j) + \sum_k \max[\rho - \mathcal{D}(I^i, C^k), 0]\}, \tag{6}$$

where I is the input features and C is their heterogeneous center. j and k mean the indexes of the positives and negatives for the i-th samples. Specifically, C^j are feature centers computed by features of positive samples of the i-th sample in the current batch. And C^k is similar but computed by negative samples. This metric learning loss essentially put the features close to their corresponding category centers and put them away from the negative centers. We add this loss on both graph features and the backbone features. Their metric learning losses are:

$$\mathcal{L}^b_{me} = \underbrace{\mathcal{L}_{hcc}(F^b, C^{F^b}, Eu)}_{\text{feature-level}} + \underbrace{\mathcal{L}_{hcc}(Y^b, C^{Y^b}, KL)}_{\text{logit-level}}, \quad \mathcal{L}^g_{me} = \underbrace{\mathcal{L}_{hcc}(Y^g, C^{Y^g}, KL)}_{\text{logit-level}}, \tag{7}$$

where the feature-level HCC loss aims to guide features embed in the euclidean (Eu) space well. And the logit-level one puts KL-divergence (KL) constraints on the classification results, further regularizing the logit distribution.

3 Graph Feature Transfer (GFT) details

In the original paper, we split the H^{2}FT into three parts in the ablation study: Graph Feature Transfer (GFT), UnBalanced Scenario simulation (UBS) and Homogeneous&Heterogeneous Graph module (H^{2}G). Here we further introduce the details of GFT. Similar to the original H^{2}FT, it can be defined as:

$$F = A \cdot X, \tag{8}$$

where X, F and A mean input features matrix, transferred features and affinity matrix respectively. X is the whole batch data consisting of N rgb and N infrared modality data (1 query with N_G galleries in inference). A is computed by the following equations:

$$A = \mathbf{D}^{-1} \cdot \mathbf{S}', \quad \mathbf{S}' = \mathcal{T}(\mathbf{S}, k), \tag{9}$$

where

$$\mathcal{S}^{i,j} = \exp \frac{cos(v(x_i), v(x_j))}{\tau}. \tag{10}$$

A is computed by the full X, which is different from our H^{2}FT. It is obvious that GFT transfers features in all batch in the training stage, suffering from the train-test modality balance gap.

4 More visualization results

In the original paper, we have given visualizations about influence of the CRI in Fig. 3. And we give more CRI visualizations about different views under the SYSU-MM01 single-shot all search mode in Fig. 2, Fig. 3, Fig. 5 and Fig. 4.

The visualizations all mean affinities with and without CRI. For each group, the first image is the sample preparing to interact and others are the top-3 similar samples of the first one. The green boxes represent correct matches, and the red boxes represent incorrect matches. The results more intuitively show the effectiveness of our method for improving affinity.

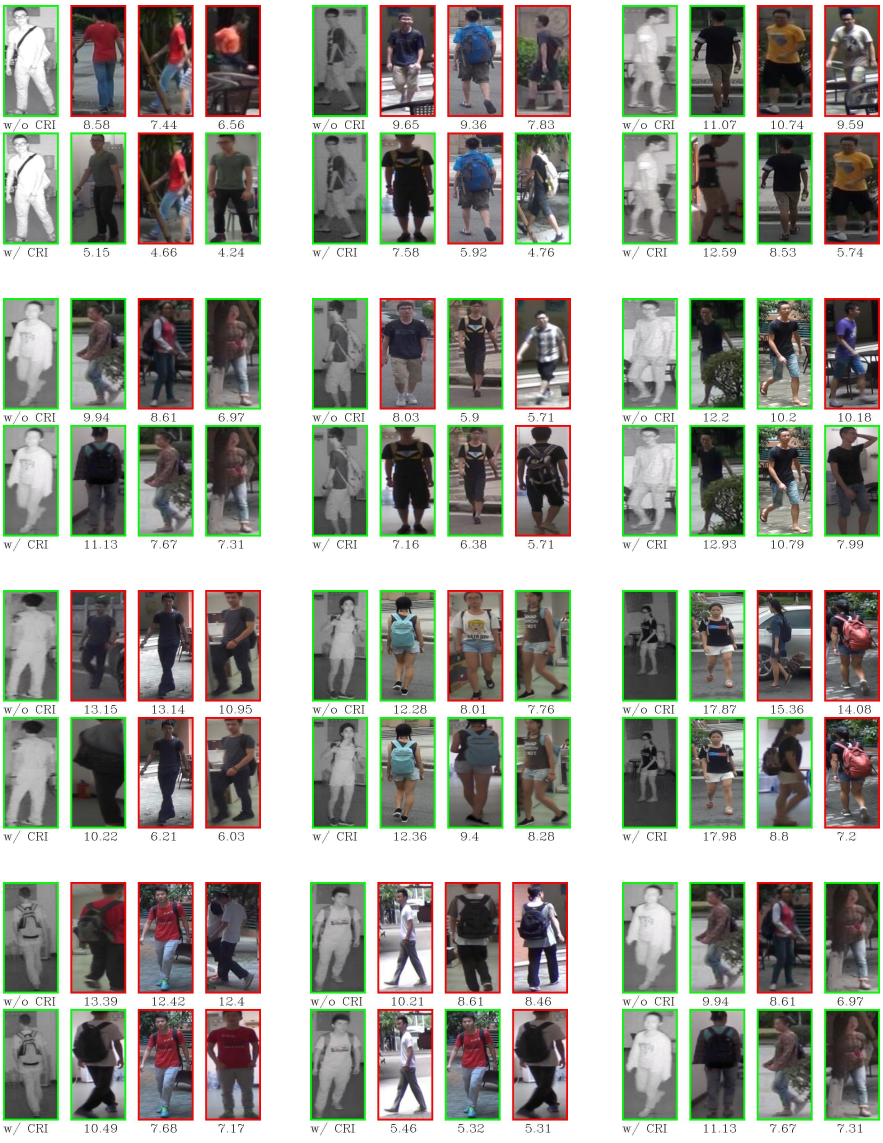

Fig. 2. CRI leads to more positive samples for graph message passing.

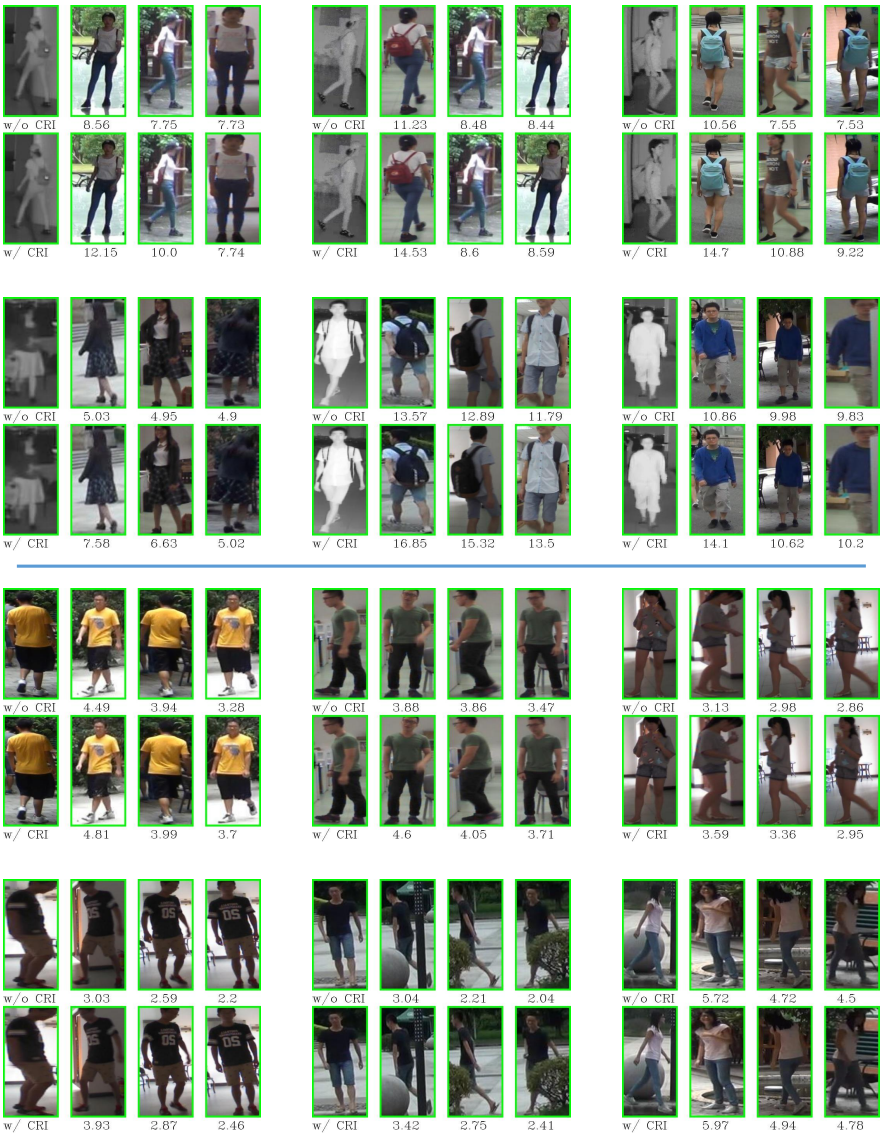

Fig. 3. CRI can enlarge positive affinities between easy sample pairs.

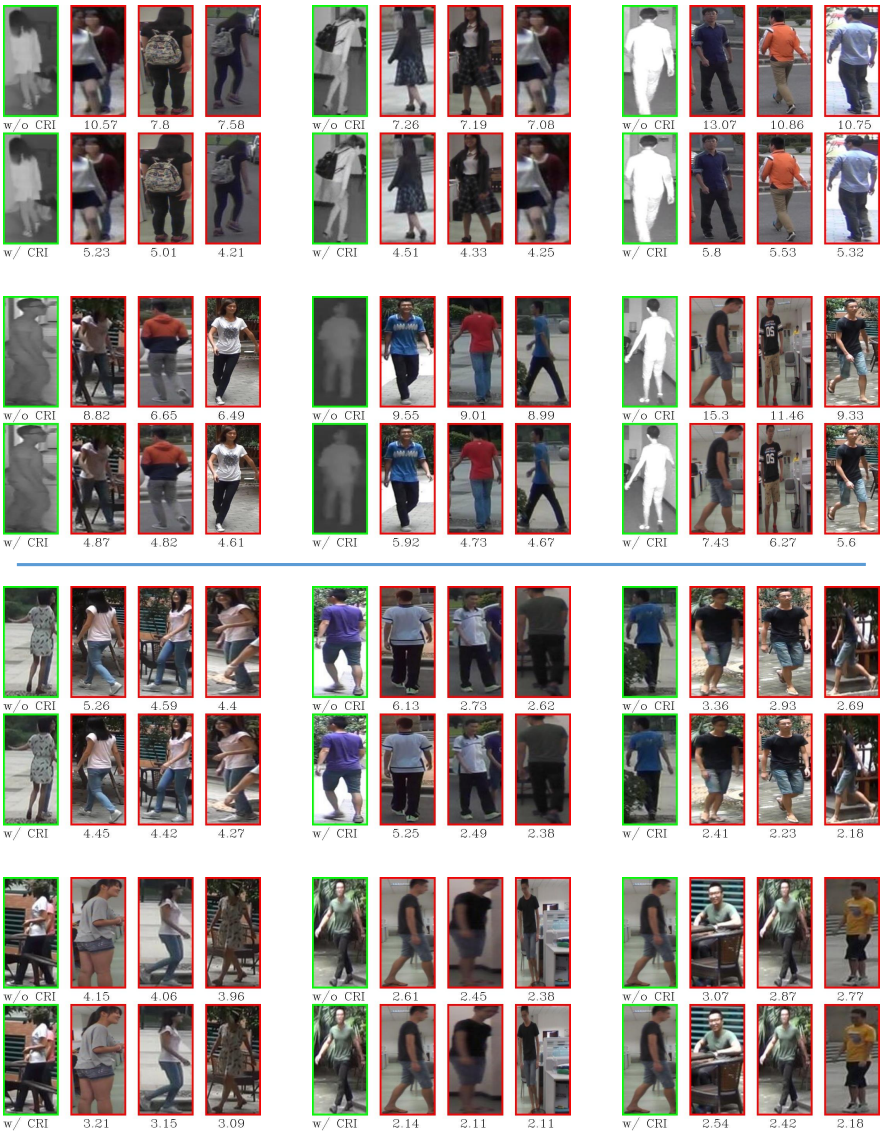

Fig. 4. CRI can suppress negative affinities between hard samples.

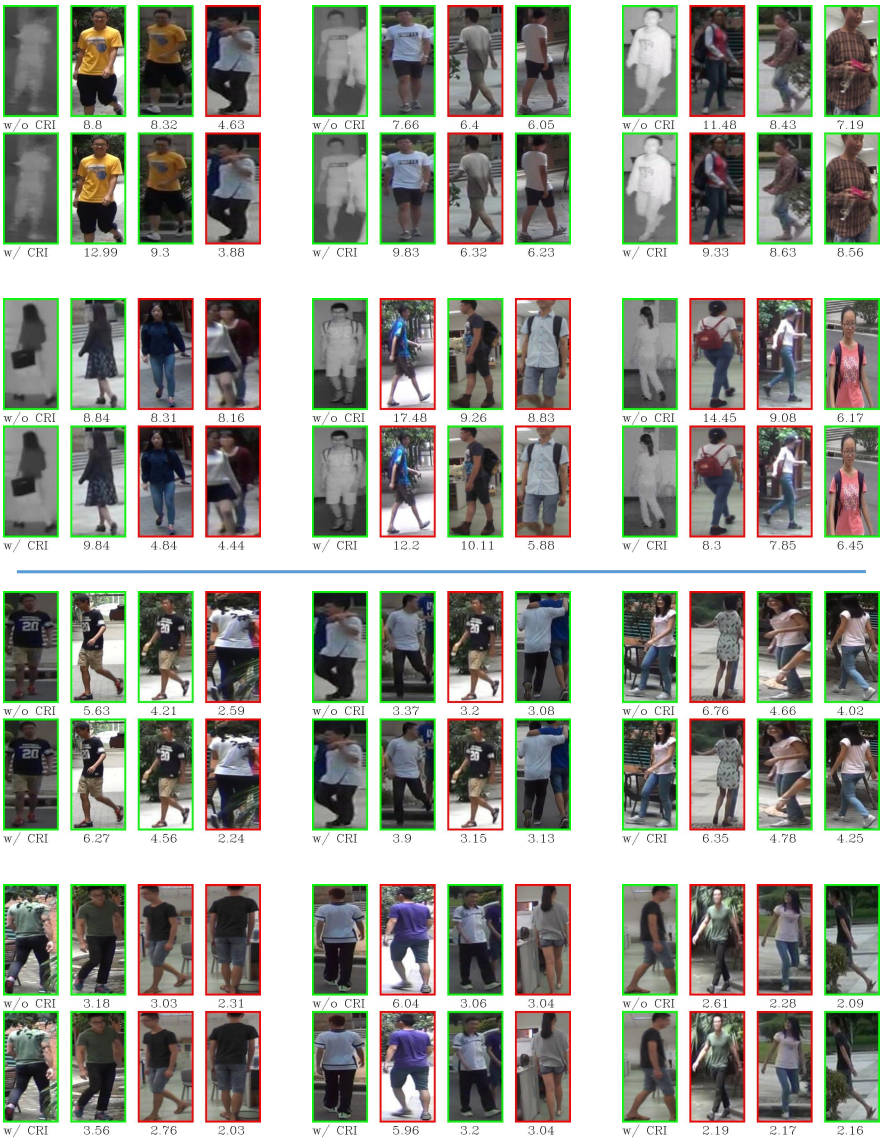

Fig. 5. CRI can enlarge positive affinities and suppress negative ones simultaneously.