Strong but Simple Baseline with Dual-Granularity Triplet Loss for Visible-Thermal Person Re-Identification

Haijun Liu, Yanxia Chai, Xiaoheng Tan, Dong Li and Xichuan Zhou

Abstract—In this letter, we propose a conceptually simple and effective dual-granularity triplet loss for visible-thermal person re-identification (VT-ReID). In general, ReID models are always trained with the sample-based triplet loss and identification loss from the fine granularity level. It is possible when a center-based loss is introduced to encourage the intra-class compactness and inter-class discrimination from the coarse granularity level. Our proposed dual-granularity triplet loss well organizes the sample-based triplet loss and center-based triplet loss in a hierarchical fine to coarse granularity manner, just with some simple configurations of typical operations, such as pooling and batch normalization. Experiments on RegDB and SYSU-MM01 datasets show that with only the global features our dual-granularity triplet loss can improve the VT-ReID performance by a significant margin. It can be a strong VT-ReID baseline to boost future research with high quality.

Index Terms—Visible-thermal person re-identification, dual-granularity triplet loss, fine to coarse granularity.

I. INTRODUCTION

VISIBLE-thermal person re-identification (VT-ReID) is a cross-modality problem and widely encountered in practical scenarios, a 24-hour intelligent surveillance system [13]. Compared to visible-visible ReID (VV-ReID), which focuses on searching a person of interest from multi-disjoint RGB cameras deployed at different locations [6], [20]. VT-ReID is a more challenge task, since person images are from different modalities. Apart from the intra-modality (intra- and inter-class) variations as existed in VV-ReID, VT-ReID additionally suffers from the large cross-modality discrepancy, arisen from the different reflective visible spectra and sensed emissivities of visible and thermal (or infrared) cameras.

Recently, an increasing number of researchers have concentrated on the VT-ReID task, achieving substantial progresses with some novel and effective modules and training strategies [7]. However, many works evaluated the effectiveness of their proposed approaches with a poor baseline, which may seriously impede the development of VT-ReID community, since the improvement of baseline model plays an important role. Therefore, in the present study, we focus on developing a strong and effective baseline for VT-ReID with some simple and typical means.

Generally, to simultaneously address the intra-modality variations and cross-modality discrepancy, different methods have been proposed, mainly focusing on the following two aspects: 1) model design and 2) metric learning.

To alleviate the extra cross-modality discrepancy, a two-stream framework is always adopted, including two modality-specific networks with independent parameters for feature extraction, and a parameter-shared network for feature embedding. In the literature, the ResNet50 [3] model is preferentially adopted as the backbone to construct the two-stream network, all the res-convolution blocks for feature extraction and some parameter-shared fully connected layers for feature embedding. In addition, the network is always trained with identification loss and triplet loss to simultaneously enlarge the inter-class distance and encourage the intra-class compactness.

In order to achieve great performance, researchers in the academia always aggregate several part (local) features [14] or leverage semantic information from pose estimation [12]. However, such approaches are not the preferable choice for industry, always bringing extra consumption. In our previous study [8], we have explored how to build the two-stream network and proposed the hetero-center based triplet loss under the part person feature learning framework. Therefore, in this letter we try to adopt some simple and typical means to improve the capability of the ReID model and only use the global features extracted by the backbone model.

This letter mainly focuses on the design of an effective baseline from two aspects. On one hand, some simple and typical means are experimentally explored to obtain the global features, including the pooling and batch normalization operations. On the other hand, the organization manner of sample-based triplet loss and center-based triplet loss are also experimentally explored to guide the network training. To summarize, our dual-granularity triplet loss (DGTL), in a hierarchical fine to coarse granularity manner, could achieve superior performance on RegDB [10] and SYSU-MM01 [16] datasets, respectively. It can be a new baseline for VT-ReID with only the global features, through a simple but effective strategy.

II. DUAL-GRANULARITY TRIPLET LOSS BASED FRAMEWORK

In this section, we introduce the framework of our proposed baseline model for VT-ReID, as depicted in Fig. 1. The framework mainly consists of two components: (1) the two-stream backbone network, and (2) the dual-granularity triplet loss module.
A. Two-stream backbone network

Based on the observation of our previous study \[8\], we empirically set the two-stream backbone network as Fig. 1. The ResNet50 [3] model is preferentially adopted as the backbone to construct the two-stream network. The shallow convolution block (stage0) and the first res-convolution block (stage1) are set as the modality-specific modules with independent parameters to learn the modality-specific 3D information from two heterogeneous modalities. Then the remaining 3 res-convolution blocks (stage2, stage3 and stage4) are set as the modality-shared module with shared parameters to learn the multi-modality shared feature maps for cross-modality re-identification by projecting those modality-specific feature maps into a modality-shared common 3D feature space.

B. Dual-granularity triplet loss module

In this module, there are three aspects should be well studied, 1) the pooling methods, 2) the batch normalization neck and 3) loss function.

1) Pooling methods: After obtaining the 3D person feature maps from the two-stream backbone network, we should firstly translate them into 1D feature vectors. We experimentally study tree kinds of the pooling methods, the global average pooling (Avg), the global max pooling (Max) and the generalized-mean pooling (GeM) \[11\].

2) Batch normalization neck: The batch normalization neck (BNNeck) \[9\] is firstly introduced in VV-ReID to address the inconsistent problem of identification and metric losses in the same embedding space. Namely, the metric loss and identification loss should process different feature vectors, before or after the batch normalization layer. However, in our framework, the BNNeck is only applied in the fine granularity branch, while in the coarse granularity branch the metric and identification losses are both applied to the features after the batch normalization layer (f_{bnf}), as shown in Fig. 1.

3) Dual-granularity triplet loss: Our previous study \[8\] only concentrates on the center-based triplet loss, which is in the coarse granularity level. Here, we simultaneously consider the sample-based triplet loss and center-based triplet loss by constructing two branches, and arrange them in a hierarchical fine to coarse granularity manner. One branch focuses on the fine granularity level with sample-based triplet loss (L_{tri}) and identification loss (L_{id}) from the coarse granularity level to obtain features f_{bnf}, following the additional red lines. During testing, the f_{bn} and f_{bnf} with L2 normalization can be adopted as the person features.
which is defined for a mini-batch \(X\), where a data point \(x^i_n\) denotes the \(n^{th}\) image feature of the \(i^{th}\) person in the batch, \([x]_+ = \max(x, 0)\) denotes the standard hinge loss, \(\|x_a - x_p\|_2\) denotes the Euclidean distance of data point \(x_a\) and \(x_p\), \(m\) is the margin.

Hetero-center triplet loss: First, in a mini-batch, the feature centers of every identity from each modality are computed, \(c^i_v = \frac{1}{K} \sum_{j=1}^{K} v^j_v\), \(c^i_t = \frac{1}{K} \sum_{j=1}^{K} t^j_t\), which is defined for a mini-batch, where \(v^j_v\) denotes the \(j^{th}\) visible image feature of the \(i^{th}\) person in the mini-batch, while \(t^j_t\) corresponds to the thermal image feature. Then, in our VT-ReID, based on the \(PK\) sampling strategy and calculated centers, we can define the hetero-center triplet loss as,

\[
L_{hc_{tri}}(C) = \sum_{i=1}^{P} \left[ mc + \|c^i_v - c^t_i\|_2 - \min_{i \neq j \in \{v,t\}} \|c^i_v - c^i_n\|_2 \right] + \sum_{i=1}^{P} \left[ mc + \|c^i_t - c^t_i\|_2 - \min_{i \neq j \in \{v,t\}} \|c^i_t - c^i_n\|_2 \right] ,
\]

where \(c^i_v\) and \(c^i_t\) are the thermal and visible feature centers of the \(i^{th}\) person, \(mc\) stands for the margin. Each identity, \(L_{hc_{tri}}\) concentrates on only one cross-modality positive pair and the mined hardest negative pair in both the intra- and inter-modality.

Finally, the dual-granularity triplet loss is,

\[
L_{all} = L_{tri}(f_p) + L_{id}(f_b) + L_{hc_{tri}}(f_{bnf}) + L_{id}(f_{bnf}),
\]

where \(f_p\), \(f_b\) and \(f_{bnf}\) are the person features as shown in Fig.4.

III. EXPERIMENTS

We evaluate the effectiveness of our proposed method for VT Re-ID tasks on two public datasets, RegDB [10] and SYSU-MM01 [16]. The implementation of our method is with the Pytorch framework. The training and testing procedures are following the official settings as done in [8], [20]. For the \(PK\) sampling strategy, we set \(P = 8, K = 4\) for the RegDB, and \(P = 6, K = 8\) for the SYSU-MM01. The pooling method is Max in both fine and coarse branches for RegDB, while Avg in fine branch and Max in coarse branch for SYSU-MM01. We set \(m = 0.3\), \(mc = 0.3\) for RegDB, and \(mc = 0.8\) for SYSU-MM01. The fusion method is element-wise sum.

A. Comparison to the state-of-the-art

In this section, our DGTL with only the global features is compared to some state-of-the-art VT-ReID methods, recently published in 2020. The results on the RegDB and SYSU-MM01 datasets are listed in Tables I and II respectively.

### Table I

| Methods | Visible to Thermal | Thermal to Visible |
|---------|--------------------|--------------------|
| CMSP [15] | IVC20 | 65.07 | 64.50 | - | - |
| HAT [21] | TIFS20 | 71.83 | 67.56 | 70.02 | 66.30 |
| MSR [17] | TIP20 | 48.43 | 48.67 | - | - |
| MACE [17] | TIP20 | 72.37 | 69.09 | 72.12 | 68.57 |
| Hi-CMD [1] | CVPR20 | 70.93 | 66.04 | - | - |
| CML [5] | MM20 | 59.81 | 60.86 | - | - |
| JSIA [13] | AAAI20 | 48.10 | 49.90 | 48.50 | 49.30 |
| XIV [4] | AAAI20 | 62.21 | 60.18 | - | - |
| DDAG [19] | ECCV20 | 69.34 | 63.46 | 68.06 | 61.08 |

B. Alation experiments

We evaluate the effectiveness of our proposed DGTL module, including three components, pooling methods, loss functions organization and the BNNecK configuration.

They show that our proposed DGTL method can achieve much better performance, especially compared to those methods with only the global features (CMSP [15], HAT [21], MSR [2], MACE [17], Hi-CMD [1], CML [5], JSIA [13] and XIV [4]), even outperforming the DDAG [19] method, which adopts the part-aggregated feature learning to refine the person features. Moreover, the results based on the \(f_{bn}\) feature, just the direct output of the ResNet50 model, also can achieve satisfactory performance, even similar to those results based on \(f_{bnf}\) feature on RegDB dataset. It demonstrates the effectiveness of our dual-granularity triplet loss module with a simple but effective strategy, which truly can be a strong VT-ReID baseline to boost further research with high quality.

Our DGTL method performs worse than HcTri [8], which is our previous study for part feature learning with much more model parameters and training tricks.
Table III and IV list the results of different pooling methods organizing in fine and coarse granularity branches, respectively. Different pooling methods truly perform differently, always with large gaps (Avg vs. Max: 70.63 vs. 80.49, rank1 on regdb dataset). Therefore, the pooling method is a key factor for constructing the VT-ReID baseline.

Table V lists the results of different triplet losses organization in the fine and coarse granularity branches. In our baseline methods (only with the fine granularity branch), the combination of $L_{tri}$ and $L_{hc,tri}$ truly can improve the VT-ReID performance. As to the dual-granularity setting, the arrangements of $L_{tri}$ and $L_{hc,tri}$ have impact on the performance. In summary, our proposed hierarchical fine to coarse granularity manner could obtain the best performance.

Table VI shows that the BNNeck [9] module only applied in the fine granularity branch is the best setting. Moreover, Fig. 2 also illustrates the effects of different fusion methods and the margin parameter in $L_{hc,tri}$.

The best performances for two datasets are with different configurations. The reason may lie in the image conditions. For RegDB, the visible and corresponding thermal images are well aligned. While for SYSU-MM01, the visible and corresponding infrared images have arbitrary poses and views.

IV. Conclusions

In this study, we propose a strong baseline for VT-ReID with a simple but effective strategy. To our best knowledge, it can achieve the best performance with only the global features of a single backbone. Our proposed DGTL method arranges the sample-based triplet loss and center-based triplet loss in a hierarchical fine to coarse granularity manner. Some simple configurations of typical operations, especially the pooling methods and batch normalization, are also explored for VT-ReID tasks. We hope that this study can promote the VT-ReID research with high quality.
REFERENCES

[1] S. Choi, S. Lee, Y. Kim, T. Kim, and C. Kim, “Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification,” in CVPR, 2020, pp. 10 257–10 266.

[2] Z. Feng, J. Lai, and X. Xie, “Learning modality-specific representations for visible-infrared person re-identification,” IEEE TIP, vol. 29, pp. 579–590, 2020.

[3] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in CVPR, 2016, pp. 770–778.

[4] D. Li, X. Wei, X. Hong, and Y. Gong, “Infrared-visible cross-modal person re-identification with an x modality,” in AAAI, 2020, pp. 4610–4617.

[5] Y. Ling, Z. Zhong, Z. Luo, P. Rota, S. Li, and N. Sebe, “Class-aware modality mix and center-guided metric learning for visible-thermal person re-identification,” in ACM MM, 2020, pp. 889–897.

[6] H. Liu and J. Cheng, “Gallery based k-reciprocal-like re-ranking for heavy cross-camera discrepancy in person re-identification,” Neurocomputing, vol. 333, pp. 64–75, 2019.

[7] H. Liu, J. Cheng, W. Wang, Y. Su, and H. Bai, “Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification,” Neurocomputing, vol. 398, pp. 11–19, 2020.

[8] H. Luo, X. Tan, and X. Zhou, “Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification,” IEEE TMM, pp. 1–1, 2020.

[9] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, “A strong baseline and batch normalization neck for deep person re-identification,” IEEE TMM, vol. 22, no. 10, pp. 2597–2609, 2020.

[10] D. Nguyen, H. Hong, K. Kim, and K. Park, “Person recognition system based on a combination of body images from visible light and thermal cameras,” Sensors, vol. 17, no. 3, p. 605, 2017.

[11] F. Radenović, G. Tolias, and O. Chum, “Fine-tuning cnn image retrieval with no human annotation,” IEEE TPAMI, vol. 41, no. 7, pp. 1655–1668, 2018.

[12] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, “Pose-driven deep convolutional model for person re-identification,” in ICCV, 2017, pp. 3980–3989.

[13] G. Wang, T. Zhang, Y. Yang, J. Cheng, J. Chang, X. Liang, and Z. Hou, “Cross-modality paired-images generation for rgb-infrared person re-identification,” in AAAI, 2020, pp. 12 144–12 151.

[14] P. Wang, Z. Zhao, F. Su, Y. Zhao, H. Wang, L. Yang, and Y. Li, “Deep multi-patch matching network for visible thermal person re-identification,” IEEE TMM, pp. 1–1, 2020.

[15] A. Wu, W.-S. Zheng, S. Gong, and J. Lai, “Rgb-ir person re-identification by cross-modality similarity preservation,” IJCV, vol. 128, pp. 1765–1785, 2020.

[16] A. Wu, W. Zheng, H. Yu, S. Gong, and J. Lai, “Rgb-infrared cross-modality person re-identification,” in ICCV, 2017, pp. 5380–5389.

[17] M. Ye, X. Lan, and Q. Leng, “Cross-modality person re-identification via modality-aware collaborative ensemble learning,” IEEE TIP, vol. 29, pp. 9387–9399, 2020.

[18] M. Ye, X. Lan, Z. Wang, and P. C. Yuen, “Bi-directional center-constrained top-ranking for visible thermal person re-identification,” IEEE TIFS, vol. 15, pp. 407–419, 2020.

[19] M. Ye, J. Shen, D. J. Crandall, L. Shao, and J. Luo, “Dynamic dual-attentive aggregation learning for visible-infrared person re-identification,” in ECCV, 2020.

[20] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, “Deep learning for person re-identification: A survey and outlook,” arXiv preprint arXiv:2001.04193, 2020.

[21] M. Ye, J. Shen, and L. Shao, “Visible-infrared person re-identification via homogeneous augmented tri-modal learning,” IEEE TIFS, vol. 16, pp. 728–739, 2021.