BRIDGING THE DISTRIBUTION GAP OF VISIBLE-INFRARED PERSON RE-IDENTIFICATION WITH MODALITY BATCH NORMALIZATION

Wenkang Li\textsuperscript{1}, Ke Qi\textsuperscript{1}, Wenbin Chen\textsuperscript{1}, Yicong Zhou\textsuperscript{2}

\textsuperscript{1}School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou, China
\textsuperscript{2}Department of Computer and Information Science, University of Macau, Taipa, Macau

ABSTRACT

Visible-infrared cross-modality person re-identification (VI-ReID), whose aim is to match person images between visible and infrared modality, is a challenging cross-modality image retrieval task. Most existing works integrate batch normalization layers into their neural network, but we found out that batch normalization layers would lead to two types of distribution gap: 1) inter-mini-batch distribution gap—the distribution gap of the same modality between each mini-batch; 2) intra-mini-batch modality distribution gap—the distribution gap of different modality within the same mini-batch. To address these problems, we propose a new batch normalization layer called Modality Batch Normalization (MBN), which normalizes each modality sub-mini-batch respectively instead of the whole mini-batch, and can reduce these distribution gap significantly. Extensive experiments show that our MBN is able to boost the performance of VI-ReID models, even with different datasets, backbones and losses.

Index Terms—Person re-identification, cross-modality, batch normalization

1. INTRODUCTION

Person re-identification is an image retrieval task, which matches person images across multiple disjoint cameras. Person re-identification plays an important role in the security field, because these cameras are usually deployed in different locations, and the results of person re-identification can help track the suspects.

In recent years, person re-identification between visible cameras has made great progress and achieved surpassing human performance on Market1501 dataset \cite{1}. However, visible cameras have poor imaging quality at night, so a lot of cameras switch to infrared mode at night. Therefore, the task of person re-identification between day and night becomes the task of person re-identification between visible and infrared. As shown in Figure \ref{fig:1}, the differences between visible images and infrared images are that infrared images are grayscale images with more noise and less details. Due to such a huge difference, the existing visible-visible person re-identification model performs poorly on the visible-infrared person re-identification task \cite{1}. In order to get better day-night person re-identification results, it is necessary to re-design models for the visible-infrared person re-identification task.

To balance the number of training samples, most existing models adopt the 2PK sampling strategy, which first randomly selects P persons, and then randomly selects K visible images and K infrared images of each selected person. It means that each mini-batch contains the same number of visible images and infrared images during the training phase. Moreover, these models integrate Batch Normalization\cite{3} layers into their neural network, so they normalize the whole mini-batch containing images of different modality. As shown in Figure \ref{fig:2} and \ref{fig:3}, we found out that this setting will lead to two types of distribution gap: 1) Inter-mini-batch distribution gap. For the same modality, we can observe that the mean and standard deviation between different mini-batches of that modality are quite different; 2) Intra-mini-batch modality distribution gap. Within the same mini-batch, we can observe that the mean and standard deviation of different modality are quite different. We believe that these distribution gaps will harm the performance of models, so we propose a new batch normalization layer called Modality Batch Normalization (MBN), which normalizes each modality sub-mini-batch respectively instead of the whole mini-batch. Comparing Figure \ref{fig:3a} with Figure \ref{fig:3b}, which applied the whole mini-batch normalization and modality sub-batch normalization respectively, we can find out that there is no...
distribution gap existing in the latter one. To demonstrate the effectiveness of our MBN, we simply replace the BN of existing models with MBN, and extensive experiments show that our MBN is able to boost the performance of VI-ReID models, even with different datasets, backbones and losses.

Our main contributions are summarized as follows:

- We found the distribution gaps caused by batch normalization and designed a new batch normalization layer called Modality Batch Normalization (MBN) to deal with this problem.

- Extensive experiments show that our MBN is able to boost the performance of VI-ReID models by simply replacing the BN with MBN.

- We establish a strong baseline for VI-ReID, which is so simple that will not conflict with most other methods, such as partial features, attention mechanisms, etc.

2. RELATED WORK

The basic solution of person re-identification is that, maps each person image into a feature embedding vector, then compute the cosine or euclidean distance between vectors as the similarity between images. For single modality person re-identification, BOT[4] establishes a strong baseline model only using global features. MGN[5] splits the output feature maps into multiple granularities and learns local features for each of them. AlignedReID[6] aligns local features between different images. ABD-Net[7] proposes a attention mechanism to enhance important areas or channels in the feature maps.

In addition to dealing with the common problems of person re-identification, visible-infrared person re-identification also needs to deal with the problems caused by modality differences. Some existing works addressed this by GAN-base methods. AligntGAN[8] aligns pixels and features at the same time. CmGAN[9] only uses adversarial learning to make the features of the two modalities indistinguishable. X Modality[10] introduces an intermediate modality. Some research is about feature learning. EDFL[11] enhances the discriminative feature learning; MSR[12] learns modality-specific representations. Some other works focus on metric learning. BDTR[13] calculates the triplet loss of intra-modality and inter-modality respectively; HPILN[14] calculates the triplet loss of inter-modality in addition to the global triplet loss; HC[15] shortens the Euclidean distance between the two modality centers. Recently, AGW[11] adopts a attention mechanism and DDAG[16] use graph neural networks to generate more useful features.
3. PROPOSED METHOD

3.1. Batch normalization and distribution gaps

Batch normalization [3] was proposed to reduce the internal covariate shift. BN first normalizes the values within the mini-batch for each channel, as illustrated in Figure 4, then linearly transforms them with learnable parameters $\gamma$ and $\beta$. Given a value $x$ $\in \mathbb{R}^{N \times C \times H \times W}$ belonging to the input feature map where $N$ is the batch size, $C$ the is channel size, $H$ is the height and $W$ is the width, BN can be expressed as:

$$\hat{x}_{n,c,h,w} = BN(x_{n,c,h,w}) = \gamma_c \bar{x}_{n,c,h,w} + \beta_c$$

(1)

$\gamma_c$ and $\beta_c$ are learnable parameters of each channel and

$$\bar{x}_{n,c,h,w} = \frac{x_{n,c,h,w} - \mu_c}{\sqrt{\sigma_c^2 + \epsilon}}$$

(2)

$\epsilon$ is a small constant value to avoid divide-by-zero, $\mu_c$ and $\sigma_c^2$ are computed by:

$$\mu_c = \frac{\sum_{n=1}^{N} \sum_{h=1}^{H} \sum_{w=1}^{W} x_{n,c,h,w}}{N \times H \times W}$$

(3)

$$\sigma_c^2 = \frac{\sum_{n=1}^{N} \sum_{h=1}^{H} \sum_{w=1}^{W} (x_{n,c,h,w} - \mu_c)^2}{N \times H \times W}$$

(4)

In the test phase, the batch size of the input may be 1, which means that computing the $\mu_c$ and $\sigma_c^2$ is useless. To deal with this, BN uses moving average $\bar{\mu}_c$ and $\bar{\sigma}_c^2$ recorded during training phase, which are computed by:

$$\mu_c^{(t+1)} = (1-\alpha) \bar{\mu}_c^{(t)} + \alpha \mu_c^{(t)}$$

(5)

$$\left(\sigma_c^{(t+1)}\right)^2 = (1-\alpha) \left(\bar{\sigma}_c^{(t)}\right)^2 + \alpha \left(\sigma_c^{(t)}\right)^2$$

(6)

$\alpha$ is the momentum factor and $t$ represents the $t^{th}$ mini-batch.

The intuition behind BN is that, the importance of each channel of the feature maps should be determined by the network, not by the input itself. So BN makes each channel distributed around zero, then learns to scale and shift each channel. However, the whole mini-batch normalization method employed by BN is not suitable for visible-infrared person re-identification, because it will lead to two types of distribution gap. As shown in Figure 5a, the mean and standard deviation of different modality sub-batches within the same mini-batch are quite different, though the whole mini-batch has already been zero mean and unit standard deviation. That is the intra-mini-batch modality distribution gap. Also shown in Figure 5a, the mean and standard deviation of the same modality between different mini-batches are quite different. That is the inter-mini-batch distribution gap. The intra-mini-batch modality distribution gap is a quite strong assumption provided by inputs, we argue that it should be determined by the network rather than the input itself, just like the intuition behind BN. What’s more, even if such a distribution gap is beneficial to the network, the inter-mini-batch distribution gap shows that it’s fluctuating. To deal with these, we propose a new batch normalization layer called Modality Batch Normalization (MBN).

3.2. Modality batch normalization

Since the whole mini-batch normalization will lead to the two types of distribution gap, we normalize each modality sub-mini-batch respectively, as illustrated in Figure 4. Assuming that $V$, $I$ includes all the visible samples and infrared samples within the mini-batch respectively, we denote $M \in \{V, I\}$.

the mean $\mu_{M,c}$ and the variance $\sigma_{M,c}^2$ of each channel belonging to each modality are computed by:

$$\mu_{M,c} = \frac{\sum_{n=1}^{N} \sum_{h=1}^{H} \sum_{w=1}^{W} x_{n,c,h,w}}{M \times H \times W}$$

(7)

$$\sigma_{M,c}^2 = \frac{\sum_{n=1}^{N} \sum_{h=1}^{H} \sum_{w=1}^{W} (x_{n,c,h,w} - \mu_{M,c})^2}{M \times H \times W}$$

(8)

So the normalized values are computed by:

$$\hat{x}_{n,c,h,w} = \begin{cases} \frac{x_{n,c,h,w} - \mu_{M,c}}{\sqrt{\sigma_{M,c}^2 + \epsilon}} & n \in V \\ \frac{x_{n,c,h,w} - \mu_{M,c}}{\sqrt{\sigma_{M,c}^2 + \epsilon}} & n \in I \end{cases}$$

(9)

We record the moving average $\bar{\mu}_{M,c}$ and $\bar{\sigma}_{M,c}^2$ for each modality:

$$\mu_{M,c}^{(t+1)} = (1-\alpha) \bar{\mu}_{M,c}^{(t)} + \alpha \mu_{M,c}^{(t)}$$

(10)

$$\left(\sigma_{M,c}^{(t+1)}\right)^2 = (1-\alpha) \left(\bar{\sigma}_{M,c}^{(t)}\right)^2 + \alpha \left(\sigma_{M,c}^{(t)}\right)^2$$

(11)

Comparing Figure 5a and Figure 5b which applied the whole mini-batch normalization and modality sub-batch normalization respectively, we can find out that there is no distribution gap existing in the latter one. The last thing to determine is whether we should share learnable affine parameters between modalities. As discussed before, if the modality distribution differences can help the network, we should use modality specific learnable parameters to make it capable of...
Taking advantage of modality differences. But if these modality differences harms, it’s hard for network to align two learnable parameters if we don’t share these learnable parameters. It is difficult to decide, so we proposed two types of MBN, which are marked as $MBN_{\text{shared}}$ and $MBN_{\text{specific}}$. The difference between the two is that the former shares learnable affine parameters between modalities, while the latter does not.

$$MBN_{\text{shared}}(x_{n,c,h,w}) = \gamma_c x_{n,c,h,w} + \beta_c$$  \hspace{1cm} (12)

$$MBN_{\text{specific}}(x_{n,c,h,w}) = \begin{cases} 
\gamma_{V,c} x_{n,c,h,w} + \beta_{V,c} & n \in V \\
\gamma_{I,c} x_{n,c,h,w} + \beta_{I,c} & n \in I 
\end{cases}$$ \hspace{1cm} (13)

3.3. Model pipeline

The overall model pipeline is shown in Figure 5. Our model pipeline is modified from BOT [4], which is strong and simple person re-identification baseline model. Comparing with the origin model, we replace all the BN with our MBN, including backbone and head; To keep simple, we use Circle Loss [17], which is a variant of softmax loss, as the loss function instead of softmax loss with triplet loss. Others are kept unchanged. We use cosine value as the similarity metric of embedding vectors.

4. EXPERIMENTS

4.1. Experiment settings

We evaluate our methods on SYSU-MM01 [2] dataset and RegDB [18] dataset. The training set of SYSU-MM01 contains 22258 visible images and 11909 infrared images from 395 IDs. The test set contains 6775 visible images and 3803 infrared images from another 96 IDs. We follow the evaluation protocol of SYSU-MM01, and report the results of all-search one-shot setting. RegDB contains 412 IDs, each ID has 10 visible images, 10 infrared images, a total of 8240 images. We follow the evaluation protocol in Ye et al. [19] for RegDB. We report the CMC and mAP metrics.

4.2. Implement details

The backbone containing MBN is initialized with ImageNet pretrained weights. The input images are resized to 320 $\times$ 128 for SYSU-MM01 and 256 $\times$ 128 for RegDB. Random erasing and random horizontal flip are adopted as data augmentation. We adopt the 2PK sampling strategy, which first randomly selects P persons, and then randomly selects K visible images and K infrared images of each selected person. We set P=6, K=8 for SYSU-MM01 and P=8, K=8 for RegDB. We use the Adam optimizer with lr=6e-4 and wd=5e-4. We warm up 2 epochs and decay the learning rate with 0.1, 0.01 at the 12th epoch and the 16th epoch respectively.
4.3. Experiment results

4.3.1. Results of Circle Loss

The results of Circle Loss are shown in Table 1. We make several observations through this: 1) If $MBN_{shared}$ or $MBN_{specific}$ is applied to the entire model, there can be a 5% increase on Rank-1 and mAP for the baseline model. 2) Applying $MBN_{shared}$ to the backbone or head alone can improve the performance of the baseline model, but applying it to the backbone alone has a greater performance improvement. 3) Applying $MBN_{specific}$ to backbone alone will reduce performance, while applying it to head alone can improve performance. 4) Mixing $MBN_{shared}$ and $MBN_{specific}$ is no better than using only one of them.

As shown in Figure 6, we plot the statistics of $|\mu_{V,c} - \mu_{I,c}|$, the absolute value of channel mean difference of different modalities, which can reflect the intra-mini-batch modality distribution gap. Compared with BN, the modality distribution gap of $MBN_{shared}$ on Backbone is much smaller, while $MBN_{specific}$ has a larger modality distribution gap due to additional affine parameters. This is why applying $MBN_{shared}$ to the backbone alone gets good results but $MBN_{specific}$ get bad results. However, the final MBN on the head, whether it is a shared or specific version, reduces the modality distribution gap to a very low level, which is why the two versions of MBN ultimately have better results.

Figure 7 shows the training curves of Circle Loss. We can see that models with MBN are much easier to fit and always get better performance during training phase.

4.3.2. Results of softmax loss with triplet loss

We also evaluate our methods with softmax loss with triplet loss, which is used by BOT [4]. As shown in Table 2, most observations are similar to Circle Loss except two: 1) $MBN_{specific}$ gets better performance than $MBN_{shared}$. 2) Applying $MBN_{specific}$ to backbone alone will boost the performance while Circle Loss won’t. As shown in Figure 8, which plots the training curves of softmax loss with triplet loss, we can see that the softmax loss curves of different models are similar, but the triplet loss curve of the $MBN_{specific}$ model drops faster than the other two models. Therefore, we believe that the reason for the better performance of $MBN_{specific}$ is that triplet loss optimizes the Euclidean distance between samples, so the modality-specific affine parameters in $MBN_{specific}$ are very helpful for optimization.

4.3.3. Results of Resnext50 backbone and RegDB dataset

We also evaluate our methods with Resnext50 [20] backbone and RegDB [18] dataset, as shown in Table 3. Due to the lack of GPU memory, for Resnext50, we set P to 5 and K to 8 in the sampling strategy. We can see that both $MBN_{shared}$ and $MBN_{specific}$ can boost the performance of model with Resnext50 backbone and $MBN_{shared}$ get

| BN type     | Visible to Infrared rank-1 | Visible to Infrared mAP | Infrared to Visible rank-1 | Infrared to Visible mAP |
|-------------|-----------------------------|--------------------------|-----------------------------|--------------------------|
| BN          | 67.3                        | 64.6                     | 65.3                        | 62.8                     |
| $MBN_{shared}$ | 67.8                      | 65.5                     | 66.2                        | 64.2                     |
| $MBN_{specific}$ | 64.7                     | 62.8                     | 63.6                        | 62.1                     |

Table 2: Results of softmax loss with triplet loss on SYSU-MM01. Backbone and head mean whether they are applied MBN. Rank-1(%) and mAP(%) are reported.

| BN type     | backbone | head | rank-1 | mAP |
|-------------|----------|------|--------|-----|
| baseline    | ×        | ☑    | 50.2   | 45.7|
| $MBN_{shared}$ | ☑        | ☑    | 51.1   | 47.0|
| $MBN_{specific}$ | ☑        | ☑    | 50.9   | 46.6|
| mixed       | ☑        | ☑    | 55.3   | 52.2|

Table 3: Results of Resnext50 [20] with Circle Loss on SYSU-MM01. BN Type is the BN layers used by the entire model. Rank-1(%) and mAP(%) are reported.

| BN type     | Visible to Infrared rank-1 | Visible to Infrared mAP | Infrared to Visible rank-1 | Infrared to Visible mAP |
|-------------|-----------------------------|--------------------------|-----------------------------|--------------------------|
| BN          | 51.7                        | 51.1                     | 53.3                        | 52.4                     |
| $MBN_{shared}$ | 53.3                     | 52.4                     | 55.3                        | 52.2                     |
| $MBN_{specific}$ | 52.4                     | 52.0                     | 56.2                        | 52.1                     |

Table 4: Results of Circle Loss on RegDB. BN Type is the BN layers used by the entire model. Rank-1(%) and mAP(%) are reported.
better performance than $MBN_{specific}$. As for the RegDB, $MBN_{shared}$ improves the performance, while $MBN_{specific}$ drops. Considering the small scale of RegDB, we think that the additional affine parameters make $MBN_{specific}$ model overfitting.

4.3.4. Comparison with state-of-the-art methods

As shown in Table 5, we compare our methods with state-of-the-art methods. The following observations can be made: 1) With the help of MBN, the Rank-1 and mAP of our model outperform most existing models except HC [15], which employs local features while ours only employ global features. 2) The Rank-10 and Rank-20 are still not as good as SOTA models. It makes sense, because our model only focuses on resolving modality differences, and don’t introduce complex methods such as attention mechanisms to deal with hard cases such as changes in person poses. Therefore, the improvement of hard cases is limited.

| method               | rank-1  | rank-10 | rank-20 | mAP    |
|----------------------|---------|---------|---------|--------|
| cmrGAN [9]           | 26.97   | 67.51   | 80.56   | 27.80  |
| eBDTR [13]           | 27.82   | 67.34   | 81.34   | 28.42  |
| EDFL [11]            | 36.94   | 85.42   | 93.22   | 40.77  |
| MSR [12]             | 37.35   | 83.40   | 93.34   | 38.11  |
| HPIILN [14]          | 41.36   | 84.78   | 94.51   | 42.95  |
| alignGAN [8]         | 42.40   | 85.00   | 93.70   | 40.70  |
| AGW [1]              | 47.50   | 84.39   | 92.14   | 47.65  |
| X-Modality [10]      | 49.92   | 89.79   | 95.96   | 50.73  |
| DDAG [16]            | 54.75   | 90.39   | 95.81   | 53.02  |
| HC [15]              | 56.96   | 91.50   | 96.82   | 54.95  |
| baseline (ours)      | 50.91   | 85.97   | 92.88   | 49.12  |
| $MBN_{shared}$ (ours)| 56.07   | 88.59   | 94.75   | 54.28  |

Table 5: Comparison with the state-of-the-arts on SYSU-MM01 dataset with all-search on-shot setting. Rank-1, Rank-10, Rank-20(%) and mAP (%) are reported.

4.4 CONCLUSION

In this paper, we propose a new batch normalization layer called modality batch normalization (MBN), which can deal with the distribution gap between different modalities. It significantly boosts the performance of VI-ReID models by simply replacing the BN with MBN. Because the MBN model is very simple, it can be used as a baseline model and be combined with other complex methods to produce a better model. We believe this finding can help researchers develop a better visible-infrared person re-identification model.

6. REFERENCES

[1] M. Ye et al., “Deep learning for person re-identification: A survey and outlook,” arXiv preprint arXiv:2001.04193, 2020.
[2] A. Wu et al., “Rgb-infrared cross-modality person re-identification,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 5380–5389.
[3] S. Ioffe et al., “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” arXiv preprint arXiv:1502.03167, 2015.
[4] H. Luo et al., “Bag of tricks and a strong baseline for deep person re-identification,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 0–0.
[5] G. Wang et al., “Learning discriminative features with multiple granularities for person re-identification,” in Proceedings of the 26th ACM international conference on Multimedia, 2018, pp. 274–282.
[6] X. Zhang et al., “Alignedreid: Surpassing human-level performance in person re-identification,” arXiv preprint arXiv:1711.08184, 2017.
[7] T. Chen et al., “Abd-net: Attentive but diverse person re-identification,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 8351–8361.
[8] G. Wang et al., “Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3623–3632.
[9] P. Dai et al., “Cross-modality person re-identification with generative adversarial training,” in IJCAI, vol. 1, 2018, p. 2.
[10] D. Li et al., “Infrared-visible cross-modal person re-identification with an x modality,” in AAAI, 2020, pp. 4610–4617.
[11] H. Liu et al., “Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification,” Neurocomputing, 2020.
[12] Z. Feng et al., “Learning modality-specific representations for visible-infrared person re-identification,” IEEE Transactions on Image Processing, vol. 29, pp. 579–590, 2019.
[13] M. Ye et al., “Bi-directional center-constrained top-ranking for visible thermal person re-identification,” IEEE Transactions on Information Forensics and Security, vol. 15, pp. 407–419, 2019.
[14] Y.-B. Zhao et al., “Hpln: a feature learning framework for cross-modality person re-identification,” IET Image Processing, vol. 13, no. 14, pp. 2897–2904, 2019.
[15] Y. Zhu et al., “Hetero-center loss for cross-modality person re-identification,” Neurocomputing, vol. 386, pp. 97–109, 2020.
[16] M. Ye et al., “Dynamic dual-attentive aggregation learning for visible-infrared person re-identification,” arXiv preprint arXiv:2007.09314, 2020.
[17] Y. Sun et al., “Circle loss: A unified perspective of pair similarity optimization,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6398–6407.
[18] D. T. Nguyen et al., “Person recognition system based on a combination of body images from visible light and thermal cameras,” Sensors, vol. 17, no. 3, p. 605, 2017.

[19] M. Ye et al., “Hierarchical discriminative learning for visible thermal person re-identification,” in AAAI, 2018.

[20] S. Xie et al., “Aggregated residual transformations for deep neural networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1492–1500.