Supplementary Material 4330
ABD-Net: Attentive but Diverse Person Re-Identification

Tianlong Chen[1], Shaojin Ding[1*], Jingyi Xie[2*], Ye Yuan[1], Wuyang Chen[1]
Yang Yang[3], Zhou Ren[4], Zhangyang Wang[1†]
[1]Texas A&M University, [2]University of Science and Technology of China
[3]Walmart Technology, [4]Wormpex AI Research
{*wiwjp619,dshj,ye.yuan,wuyang.chen,atlaswang*}*@tamu.edu*
hsfzxjy@mail.ustc.edu.cn, *yang.yang2@walmart.com*, *renzhou200622@bianlifeng.com*

https://github.com/TAMU-VITA/ABD-Net

1. Detailed Structure of ABD-Net

1.1. Channel Attention Module

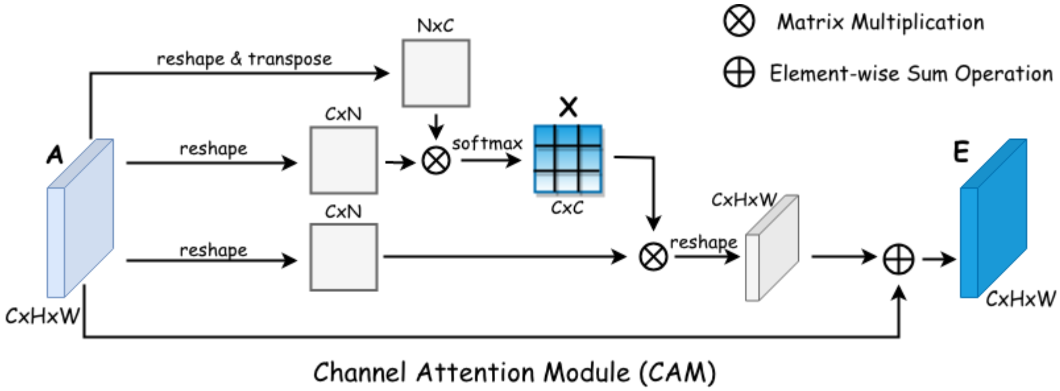

Figure 1. Channel Attention Module (CAM)

The detailed structure of CAM is illustrated in Fig. 1. Given the input feature maps $\mathbf{A} \in \mathbb{R}^{C\times H\times W}$, where C is the channel number and $H \times W$ the feature map size, we compute the affinity matrix $\mathbf{X} \in \mathbb{R}^{C\times C}$, as shown in equation (1).

$$x_{ij} = \frac{exp(A_i \cdot A_j)}{\sum_{j=1}^{C} exp(A_i \cdot A_j)}, \; i,j \in \{1,\cdots,C\} \tag{1}$$

where C is the total number of channels, and x_{ij} represents the impact of channel i on channel j. The final output feature maps **E** are calculated by the equation (2):

$$E_i = \gamma \sum_{j=1}^{C} (x_{ij}A_j) + A_i, \; i \in \{1,\cdots,C\} \tag{2}$$

γ is a hyperparameter to adjust the influence of CAM. In our experiment, $C = 1024, H = 24, W = 8$.

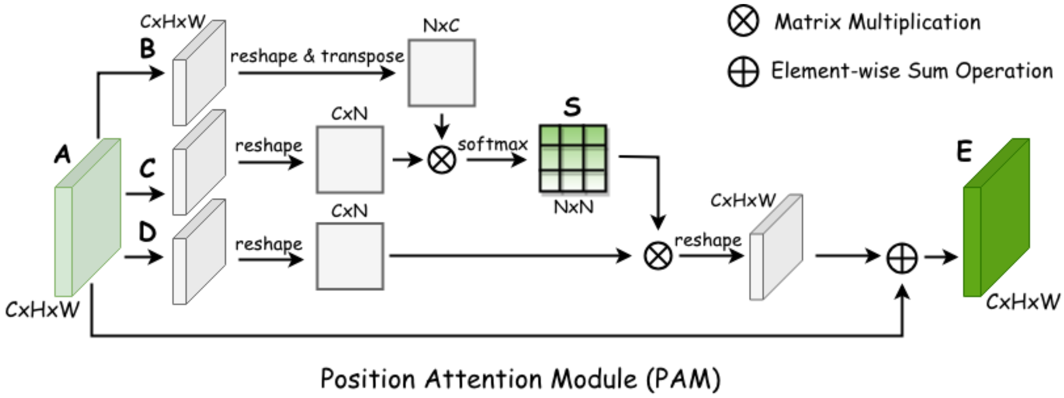

Figure 2. Position Attention Module (PAM)

1.2. Position Attention Module

The detailed structure of the Position Attention Module (PAM) is illustrated in Fig.2. The input feature maps $\mathbf{A} \in \mathbb{R}^{C\times H\times W}$ are first fed into convolution layers[1] to produce feature maps **B, C, D** $\in \mathbb{R}^{C\times H\times W}$. Then we compute the pixel affinity matrix $\mathbf{S} \in \mathbb{R}^{N\times N}$, where $N = H \times W$, as shown in equation (3).

$$s_{ij} = \frac{exp(B_i \cdot C_j)}{\sum_{j=1}^{N} exp(B_i \cdot C_j)},\ i, j \in \{1, \cdots, N\} \tag{3}$$

where s_{ij} represents the correlation between pixel i and pixel j. Finally, the output feature maps **E** are generated by averaging over the attention masks and the original feature map, as shown in equation (4):

$$E_i = \alpha \sum_{j=1}^{N} (s_{ij} D_j) + A_i,\ i \in \{1, \cdots, N\} \tag{4}$$

α is a hyperparameter to adjust the influence of CAM. In our experiment, $C = 1024, H = 24, W = 8$.

2. Detailed Implementation of ABD-Net

2.1. Training Procedure

All our ablation study and the final result are trained using the following process, which is known as the two-step transfer learning algorithm.

1. The weights of the backbone network are frozen and only reduction layers, classifier layers and all modules (PAM, CAM, O.F. and O.W.) are trained for 10 epochs. In this stage, only cross entropy loss and triplet loss are applied.

2. All layers are open for training for 20 epochs. In this stage, cross entropy loss, triplet loss, and O.W. penalty are applied.

3. All the four terms (cross entropy loss, triplet loss, O.W. penalty, and O.F. penalty) in loss are applied, and the model are trained for another 60 epochs.

We use Adam optimizer to finetune our model, with the base learning rate initialized as 3×10^{-4}, then decayed to 3×10^{-5} after 30 epochs (i.e., at the end of Stage 2), and further decayed to 3×10^{-6} after 50 epochs (i.e., 20 epochs from the start of Stage 3). O.F. penalty should be applied after the first time base learning rate is decayed, but not necessarily at the very beginning of Stage 3. In practice, it's also acceptable to delay the timing for around 5 epochs, which may reduce the training time to some extent.

[1] The convolution layer contains batch normalization and ReLU activation.

We set $\beta_{tr} = 10^{-1}$, $\beta_{OF} = 10^{-6}$ and $\beta_{OW} = 10^{-3}$, and the margin parameter for triplet loss $\alpha = 1.2$ in the final loss function (5).

$$L = L_{xent} + \beta_{tr} L_{triplet} + \beta_{O.F.} L_{O.F.} + \beta_{O.W.} L_{O.W.} \tag{5}$$

β_{tr} and α are set empirically, while the adoption of β_{OF} and β_{OW} is determined by grid search method. There are two reduction modules in our model, each with a dropout layer. We increased the parameter p in dropout layers as:

$$p = min(0.5, 0.2 + \lfloor epoch/10 \rfloor \times 0.1)$$

where $epoch$ is the epoch number from the beginning of Stage 2. During Stage 1, p is fixed as 0.2. The above parameter setting is shared across all experiments. Our network is trained using 2 Tesla P100 GPU with a batch size of 64. Each identity contains 4 instance images, and there are 16 identities per batch. Such setting is important for Triplet loss, but not necessary if only cross entropy loss is used.

During training, the input images are re-sized to 384×128 and then augmented by random horizontal flip, normalization, and random erasing. In ablation study, random erasing is used if explicitly stated. During testing, the images are re-sized to 384×128 and augmented only with normalization. Both the original and the horizontally flipped images are fed into the model, and the final feature embedding is the average of the two outputs.

3. Additional Visualizations of ABD-Net

We did additional visualizations of attention maps, correlation matrix and qualitative re-ID results from Baseline (XE), Baseline (XE) + PAM + CAM and ABD-Net (XE)[2], as shown in Fig. 3, Fig. 4, Fig. 5 and Fig. 6.

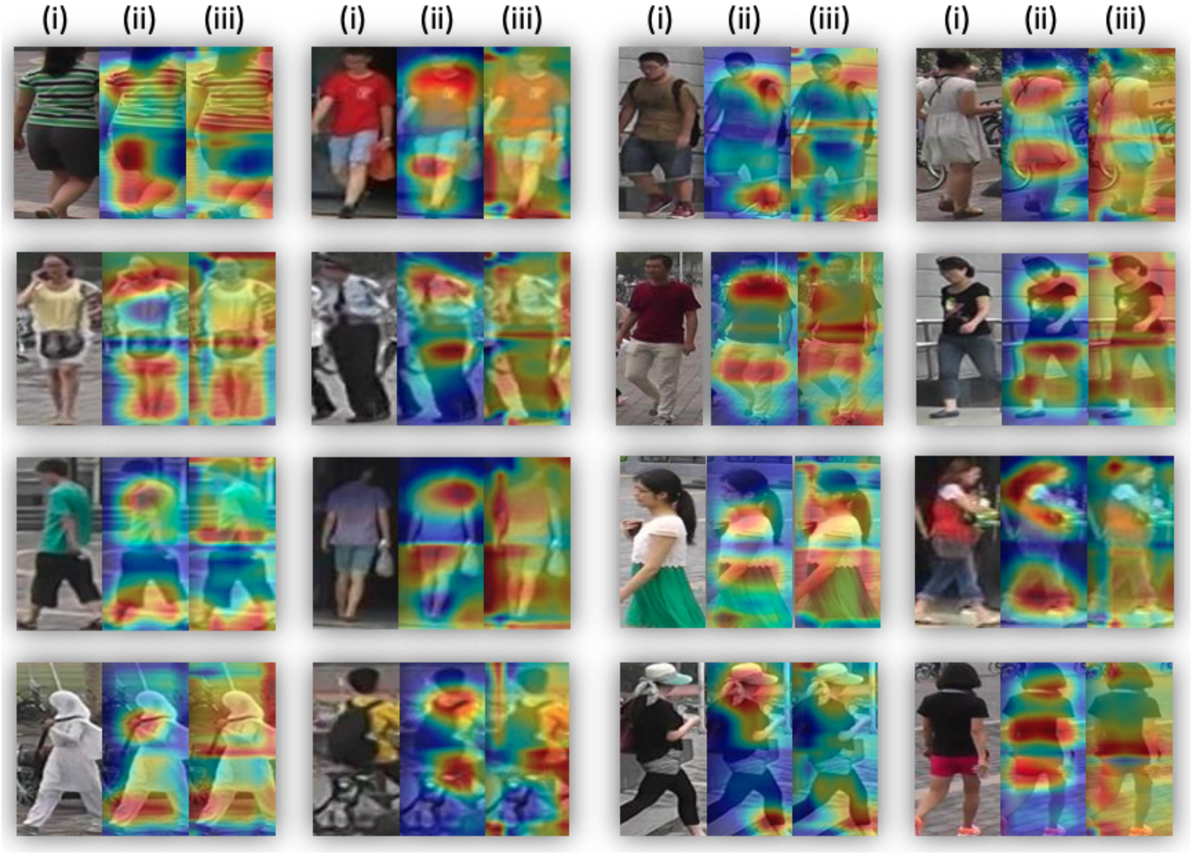

Figure 3. Visualizations of attention maps with RAM method. (i) **Testing images**; (ii) Attentive feature maps from Baseline (XE) + PAM + CAM; (iii) Attentive but diverse feature maps from ABD-Net (XE).

[2] Here we did not use our best model ABD-Net, since we hope to ensure the fair comparison of the three methods, using the same XE loss.

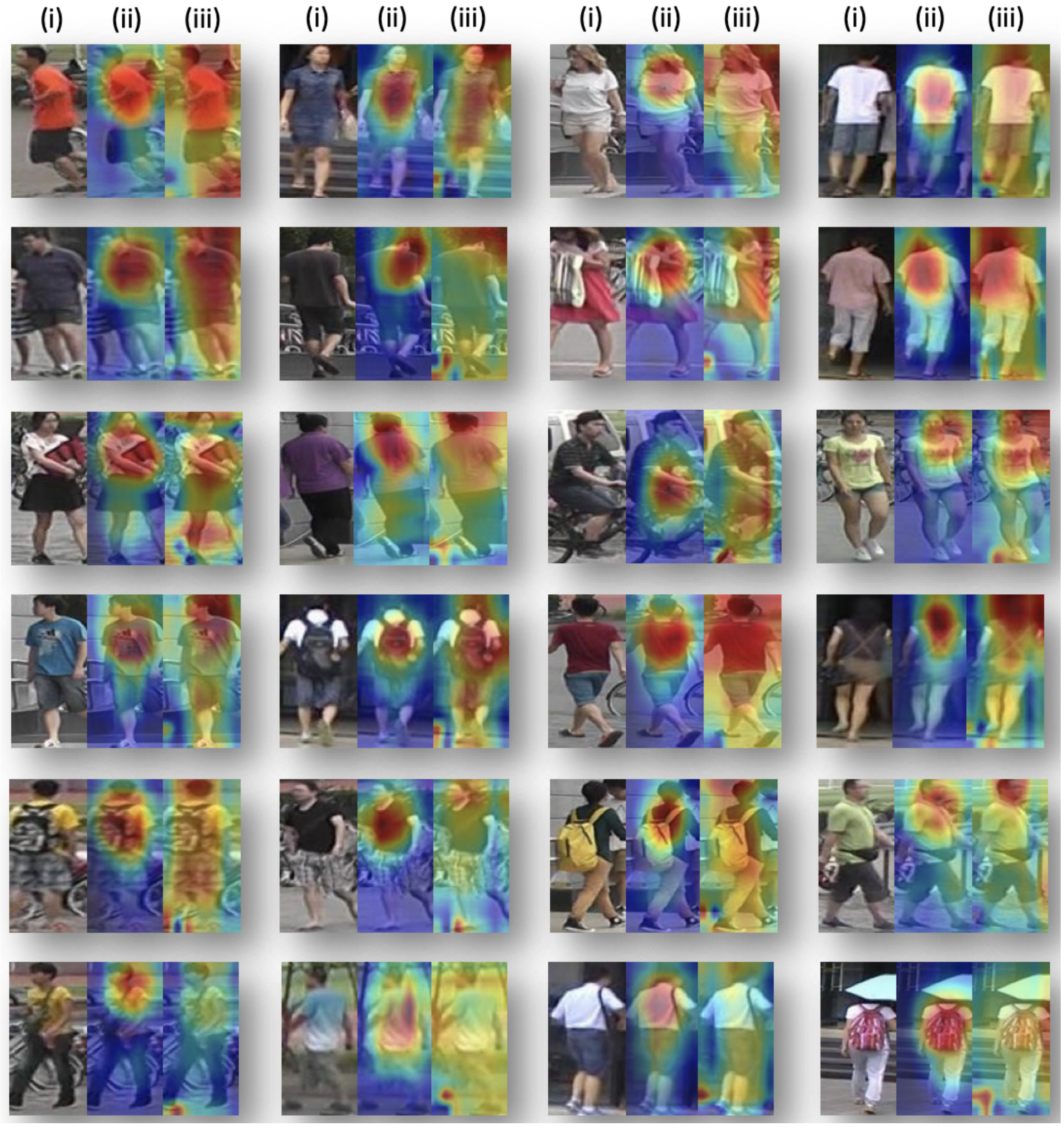

Figure 4. Visualizations of attention maps with Grad-CAM method. (i) Original images; (ii) Attentive feature maps from Baseline (XE) + PAM + CAM; (iii) Attentive but diverse feature maps from ABD-Net (XE).

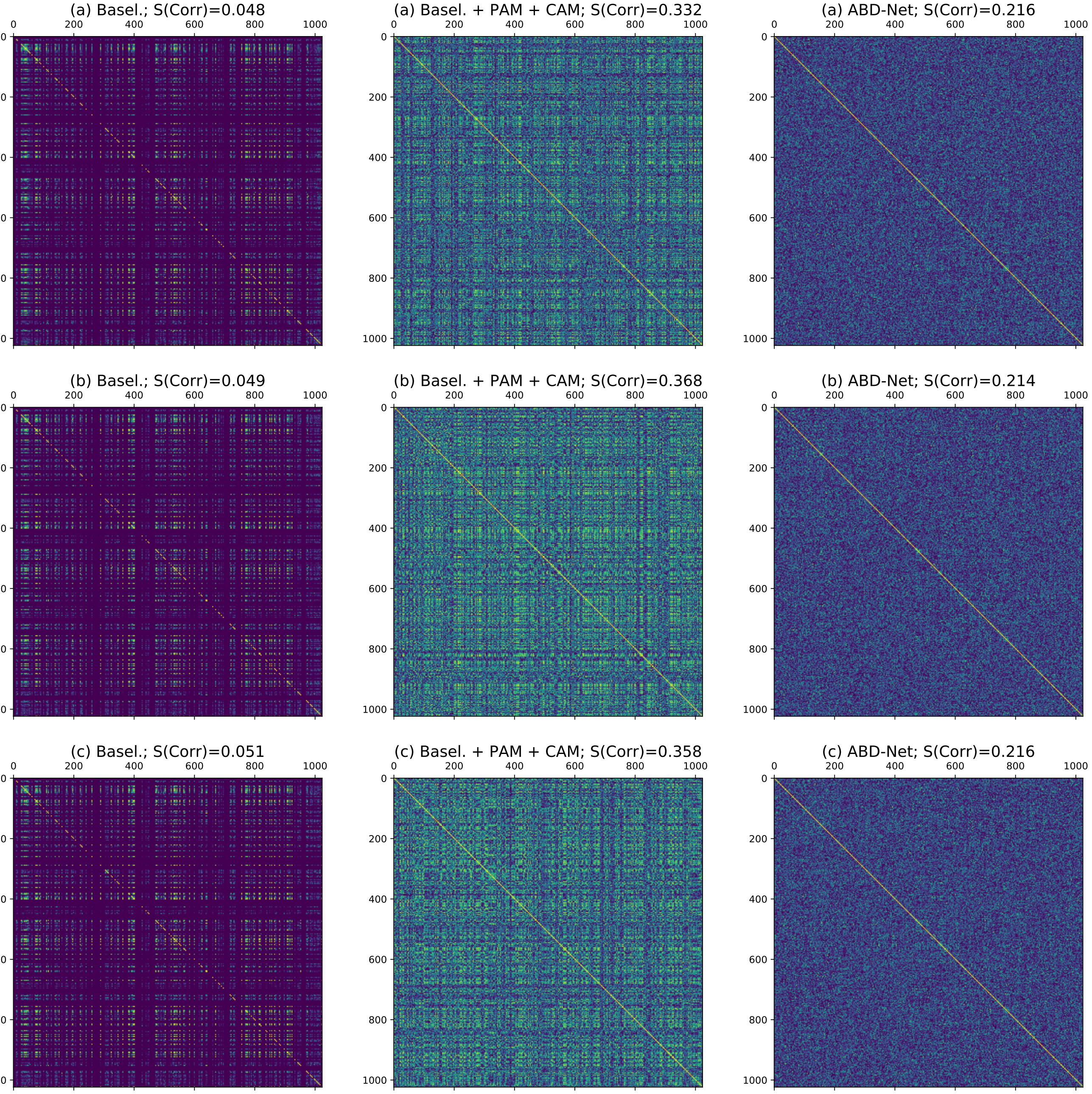

Figure 5. Visualizations of correlation matrix between channels from Baseline (XE), Baseline (XE) + PAM + CAM and ABD-Net (XE) of three random picked testing images. Brighter color indicates larger correlation. S(**Corr**) is the average of all correlation coefficients in the matrix.

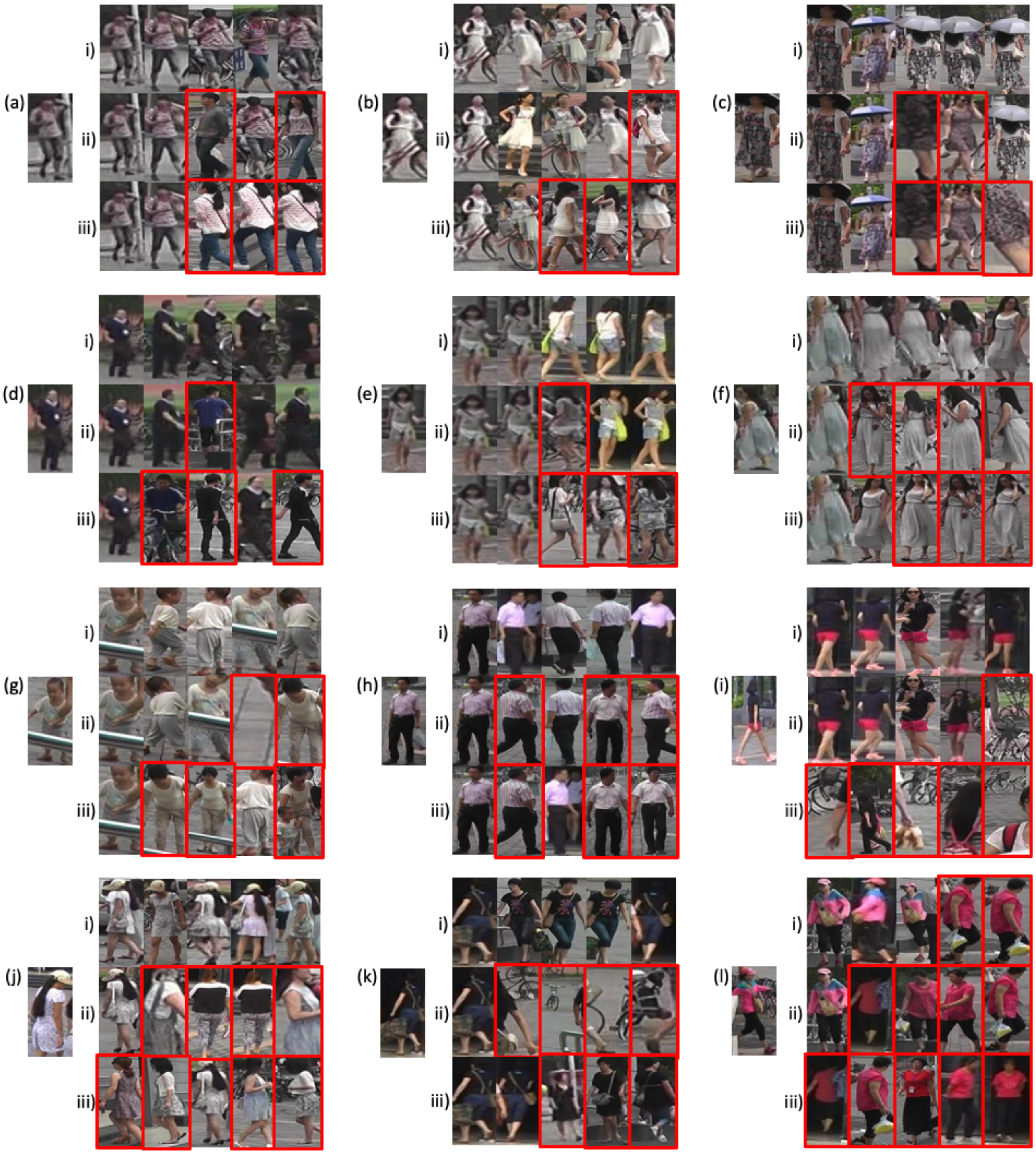

Figure 6. Twelve Re-ID examples of ABD-Net (XE), Baseline (XE) + PAM + CAM and Baseline (XE) on Market-1501. Left: query image. Right: i): top-5 results of ABD-Net (XE). ii): top-5 results of Baseline (XE) + PAM + CAM. iii): top-5 results of Baseline (XE). Images in red boxes are negative results.