Supplementary Material

Diversity Regularized Spatiotemporal Attention for Video-based Person Re-identification

3.2.1 Feature Enhancement

After running multi-region spatial attention, each frame is represented by K spatially gated features. We then pool these features across time to enhance their representation. Recall that $\mathbf{x}_{n,k} \in \mathbb{R}^D$ is the feature vector of the k^{th} spatial component from the n^{th} frame. The matrix $\mathbf{X}_k = [\mathbf{x}_{1,k}, \ldots, \mathbf{x}_{N,k}], \mathbf{X}_k \in \mathbb{R}^{D \times N}$ is defined as the set of spatial gated visual features generated by the k^{th} attention model over all frames. Because frames with similar spatial gated features tend to represent multiple observations of the same patch of pixel values, we generate a robust feature representation of each component at each frame by pooling information from the same component at other frames. For each feature extracted at frame n, we pool information from auxiliary frames by assigning a weight to each of them, which depends on the similarity of the spatial gated features as well as the temporal distance between the two frames.

We define a per-component "feature similarity" matrix $\mathbf{\Phi}_k \in \mathbb{R}^{N \times N}$ as the inner product of feature vectors

$$\mathbf{\Phi}_k = (\mathbf{X}_k)^\top \mathbf{X}_k. \tag{1}$$

Additionally, we define a "temporal similarity" matrix $\Psi \in \mathbb{R}^{N \times N}$

$$\mathbf{\Psi} = \mathbf{W} \exp\left(-\frac{|i-j|}{\sigma}\right) + \mathbf{b}, \tag{2}$$

where $|i-j|$ is a matrix that represents relative temporal distance between any frame i and frame j, and $\mathbf{W} \in \mathbb{R}^{N \times N}$ and $\mathbf{b} \in \mathbb{R}^N$ encodes the positional information.

The overall similarity scores across all frames in the video sequence of the k^{th} spatial component are represented as follows.

$$\mathbf{C}_k = \text{softmax}'(\mathbf{\Phi}_k + \mathbf{\Psi}), \tag{3}$$

where $\mathbf{\Phi}_k$ and $\mathbf{\Psi}$ formulate the appearance and location similarity respectively. Softmax$'$ is the Softmax function over each row of $(\mathbf{\Phi}_k + \mathbf{\Psi})$. Elements in the i^{th} row of $\mathbf{C}_k$ describe the contribution probabilities of all frames in the input sequence to frame i. The enhanced feature representation is written as a residual connection,

$$\widehat{\mathbf{X}}_k = FCN(\mathbf{X}_k \mathbf{C}_k) + \mathbf{X}_k, \tag{4}$$

where FCN are a linear transformations. $\widehat{\mathbf{x}}_{n,k} \in \mathbb{R}^D$ is the n^{th} element of $\widehat{\mathbf{X}}_k$ which corresponds to Eq. (4) in the paper ($\widehat{\mathbf{x}}_{n,k} = E(\mathbf{x}_{n,k})$).

4.5. More Experimental Results

Fig.1 illustrates more examples of corresponding receptive fields generated by multiple spatial attention models. Fig.2 is the visualization results of temporal attention. The temporal attention model learns the importance of each frame based on the merits of different salient regions. Our temporal attentions assign low weights to occluded or background parts and high weights to correctly detected parts which confirms the effectiveness of the model.

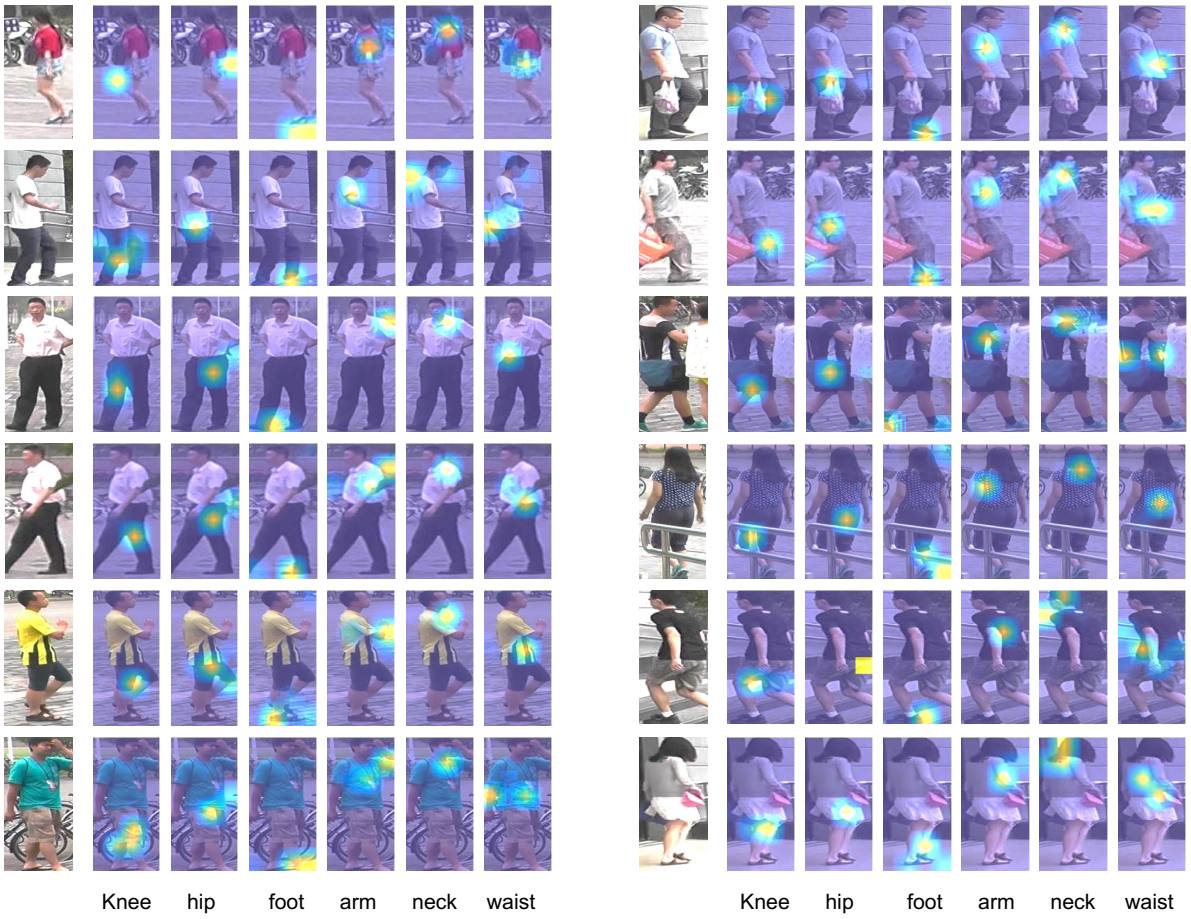

Figure 1. Example images and corresponding receptive fields for our diverse spatial attention models when $K = 6$. Our methodology discovers distinctive image regions which are useful for re-identification. The attention models primarily focus on foreground regions and generally correspond to specifc body parts. Our interpretation of each is indicated at the bottom of each column.

Figure 2. Visualization results. The numbers under images indicate the temporal attention weights assigned to each frame. Our temporal attentions assign low weights to occluded or background parts and high weights to correctly detected parts.