Automatic Adaptation of Person Association for Multiview Tracking in Group Activities

Minh Vo[1], Ersin Yumer[2], Kalyan Sunkavalli[3], Sunil Hadap[3],
Yaser Sheikh[1], and Srinivasa G. Narasimhan[1]

[1]Carnegie Mellon University, [2]Argo AI, [3]Adobe Research

1 More Analysis of the Descriptor Adaptation

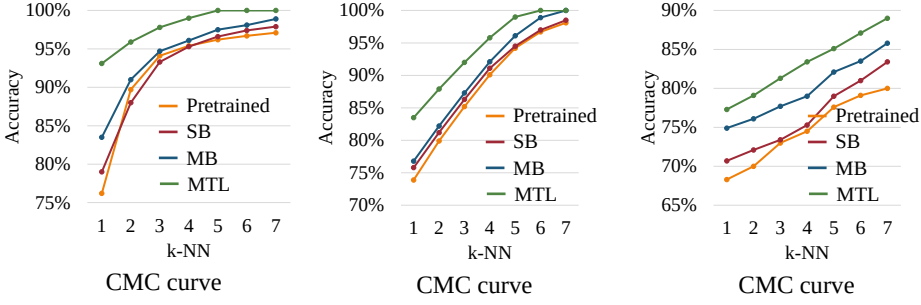


Fig. 1: The CMC for the Chasing (left), Tagging (middle), and Halloween (right) scene at different stage of our algorithm. Our method outperforms the pretrained model at every stages.

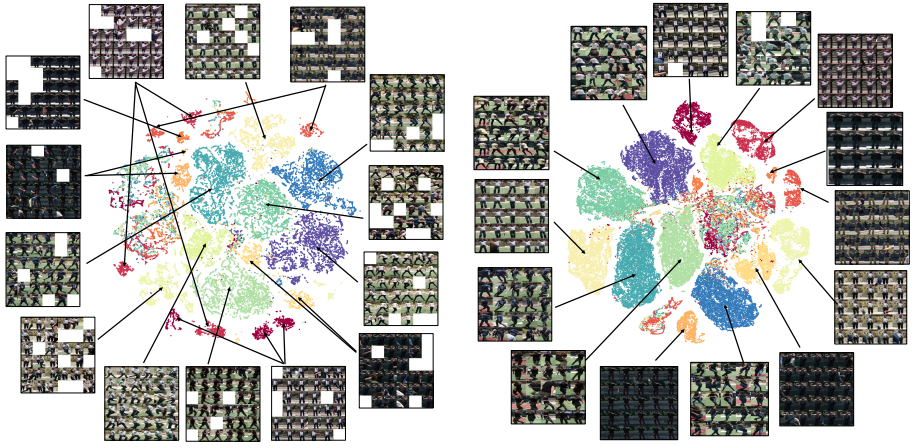
Figure 1 shows the Cumulative Matching Characteristic (CMC) for all sequences: Chasing, Tagging, and Halloween. There are clear improvement over the pretrained model as more sophisticated stages of your algorithm is applied.

Figure 2 shows the t-SNE 2D embedding of the descriptor extracted using multitask learning for the Tagging sequence. While the descriptors extracted from the pretrained model are scatter, our descriptor groups images of the same person from all views and time instances into cleanly separated clusters.

Figure 3 and Figure 4 shows the t-SNE 2D embeddings for the Halloween sequence using the pretrained model and our Multitask Learning approach. Despite being a very complex scene with high number of people, our proposed algorithm shows better discrimination and the same person is better grouped into a single cluster.

2 Pretrained Model: Pose-insensitive Person Descriptor

Our goal is to learn a generic appearance descriptor extractor of a person image that has similar output for images of the same person and dissimilar outputs for



(a) Pretrained (b) MTL

Fig. 2: t-SNE visualization of the person descriptor extracted using a pretrained model and our multitask learning (MTL) for the Tagging sequence. The pretrained descriptors are scatter. Our descriptor groups images of the same person from all views and time instances into cleanly separated clusters.

different people regardless of the viewing direction and pose deformation. One straight forward way to make the descriptor pose-insensitive is to rectify the input image into a canonical frame [1]. However, such rectification is problematic due to 2D warping artifact and wrong pose detection. Instead, we augment the RGB image with the heatmaps of keypoints and their part affinity fields provided by CPM model [2] (see Figure 5). This representation avoids the viewing direction quantization in rectifying the body parts [3, 4] and takes the detection confidence into account to down weight possible pose detection failures.

Inspired by the ability to learn multi-scale features and model compactness of the Inception architecture [5], our customized CNN model passes an input of size 288x112x59 through six convolution layers, six Inception modules, and three fully connected layers. The person descriptor is extracted at $fc8$ layer. The detailed structure are listed in Table 1. The Batch Normalization layers, which accelerate the convergence process and avoid manually tweaking the initialization of weights and biases, are employed before each ReLU layer[6].

We train this model on a combination of 18 different reID datasets. Each dataset was collected with very different locations with various camera setups and image resolution. CUHK02, CUHK03, DUKE-MTMC, MARS were captured on campus, where many students wear backpacks. PRID contains pedestrians in street views, where crosswalks appear frequently in the dataset. VIPeR images has significant illumination variation across different camera views. iLIDS was captured at the airport with many people dragging the luggages. The CMU dataset, captured in the CMU Panoptics studio, contains strong viewpoint and

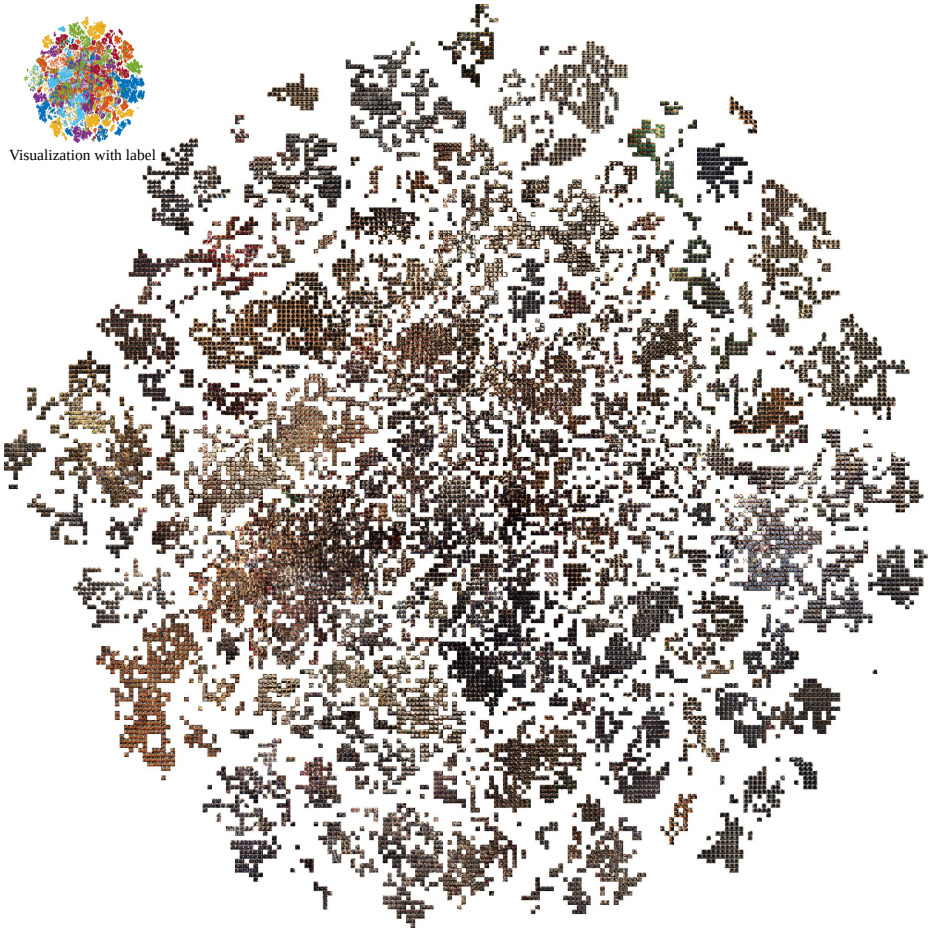


Fig. 3: tSNE visualization of the person descriptor extracted using the pretrained model for the Halloween scene.

pose variations. ETH was captured from a single moving camera in street views for multi target tracking purpose. Notably, CUHK04, contained both images captured around an urban city and movie snapshots, has rich variations of viewpoints, lighting, background conditions. This dataset has the largest number of people identity but with very few views for each of them (on average 3 images). The statistics for each of these dataset is summarized in Table 2 The combined dataset has approximately 200k images of nearly 16k identities with strong variance in background conditions, illumination, viewpoints, and people races. We regularize training by randomly switching off 50% of the neurons in the fc7 layer during training. We employ the standard softmax loss and train the model from scratch using Stochastic Gradient Descent (SGD) with the initial learning rate set to 0.1 and decreased by 8% for every 4 epochs until it reaches 0.0001.

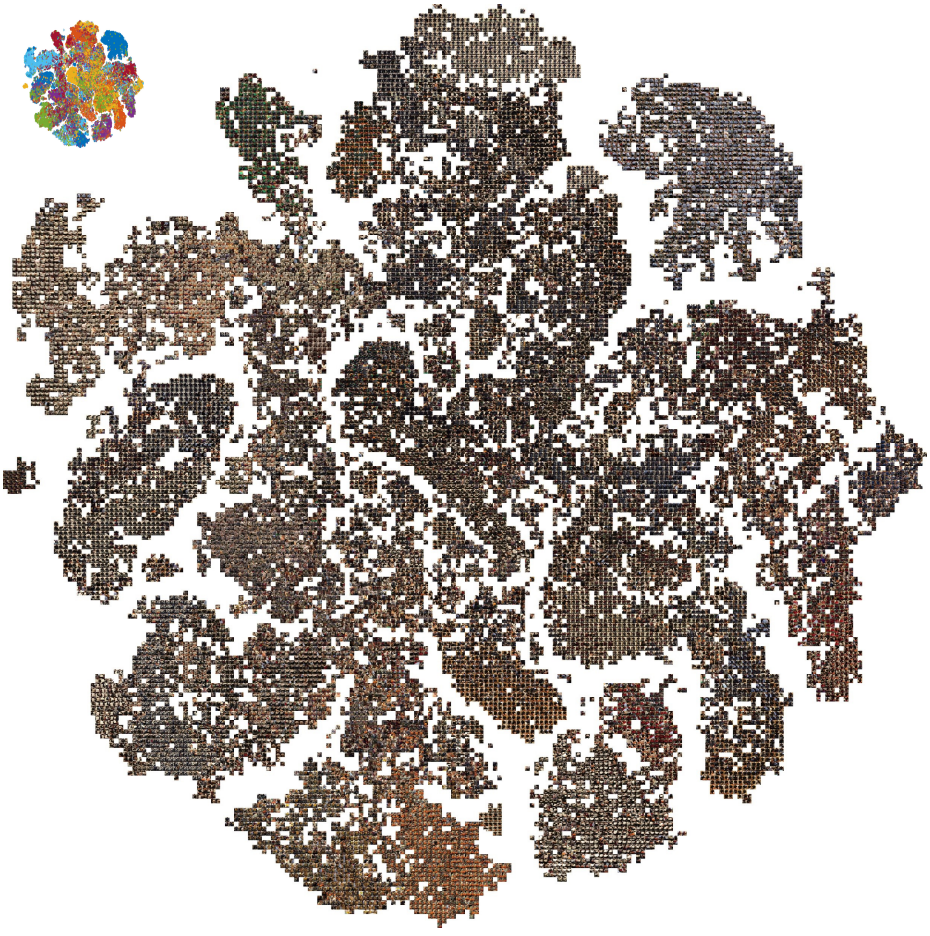


Fig. 4: tSNE visualization of the person descriptor extracted using our multitask learning (MTL) for the Halloween scene.

2.1 Analysis of Pose-Insensitive Person Descriptor

Comparison with state of the art: Table 3 shows the comparison between our approach and the state of art methods for the top-1 matches on six commonly used datasets. For video dataset such as MARS, most methods compute the averaging distance of the learned feature descriptor over all pairs of time instances of the tracklets to match between trajectories. We perform per-frame matching, which is more challenging. Our approach outperforms most other methods by a margin except for ViPER, which is a small dataset with strong variations in viewpoint, image quality, and lighting condition. Since the total number of images from PRID, iLIDS, ViPER, and 3dPES comprises less than 5% of the training images, their appearance statistics is likely to be dominated by larger datasets.

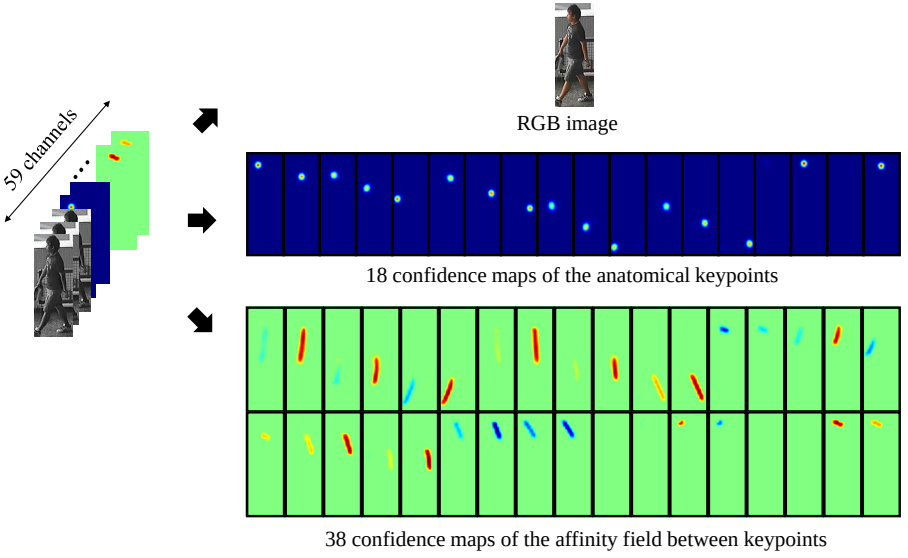


Fig. 5: The input to our CNN is a 59-channel feature maps, consisting of the color image, the feature maps of the 18 anatomical keypoints and their affinity fields computed by CPM.

This potentially explains for their lower accuracies compared to CUHK03 and MARS.

Sensitivity to viewpoint and pose variations: We analyze the pose and viewpoint insensitive properties using the data collected from the CMU Panoptic. None of them are the same people as in the CMU dataset used in the training. Figure 6 shows the t-SNE visualization [26] of the descriptor for all the images. Despite the similar appearance of multiple people, the images of the same person are clustered together. This provides a strong evidence of the pose and viewpoint insensitivity of our descriptor.

Effectiveness of the pose heatmap: We perform an ablative analysis of how different heatmap categorie affects the matching accuracy. Table 4 summaries the results. Augmenting the color images with the CPM heatmaps improves the accuracy, among which ViPER is boosted by 10.1%. Using both the keypoints and part affinity field heatmaps gives the best accuracy, albeit modest improvement over keypoints or affinity field heatmaps alone.

3 Human Mesh Model Fitting

Given the 3D human skeleton, we fit the SMPL mesh model [27] to the skeleton. Let $\mathbf{J}(\beta, \Phi, \gamma)$ be the anatomical joints regressed from the SMPL model. $\mathbf{J}(\beta, \Phi, \gamma)$ has 18 joints, defined in correspondence with the CPM joints[2], γ is

name	patch size/ stride	output side	#1×1	#3X3 reduce	#3x3	double #3X3 reduce	double #3X3	pool+proj
input		59x288x112						
conv 0 – conv 5	3x3/2	32x72x28						
inc 1a		256x72x28	64	64	64	64	64	avg+64
inc 1b	stride 2	384x36x14	64	64	64	64	64	max+identity
inc 2a		512x36x14	128	128	128	128	128	avg+128
inc 2b	stride 2	768x18x7	128	128	128	128	128	max+identity
inc 3a		1024x18x7	256	256	256	256	256	avg+256
inc 3b	stride 2	1536x9x4	256	256	256	256	256	max+identity
fc7		256						
fc8		256						
fc9		M						

Table 1: The structure of our CNN model for person ReID. This model is inspired by the Inception architecture, known for its efficiency and expressiveness.

the global 3D translation vector of the root joint, the shape parameter β is a vector of 10 numbers, and the pose parameters Φ is defined by a skeleton rig with 23 joints representing the axis-angle representation of the relative rotation between body parts. We fit the SMPL model to the 3D location of the joint K by minimizing the following cost:

$$E_{SMPL}(\beta, \Phi, \gamma) = E_d + E_p, \tag{1}$$

where E_d is the 3D fitting cost and E_p is the pose prior cost.

3D fitting cost: This cost function penalizes the Euclidean distance between the estimated joints $\mathbf{K}$ and the joints regressed from the SMPL model J:

$$E_d(\beta, \Phi, \gamma) = \sum_{t=1}^{F} \sum_{p=1}^{18} \rho\left(\frac{J^p(\beta, \Phi(t), \gamma(t)) - K^p(t)}{\sigma_{3D}} \right), \tag{2}$$

where σ_{3D} is the expected variation in triangulating the 3D joint $K^p(t)$, and ρ is the Huber estimator.

Pose prior cost: This cost function favors common poses over rare ones:

$$E_p(\Phi) = \sum_{t=1}^{F} \Phi(t)' A \Phi(t), \tag{3}$$

where A is the multivariate Gaussian normal distribution learned from the CMU motion capture database[28].

We initialize (β, Φ, γ) by rigidly aligned the torso's joint of the SMPL model $\mathbf{J}(\beta, \Phi, \gamma)$ to the estimated torso's points K and optimize 1 using the LevenbergMarquardt method implemented in the Ceres Solver package [29].

	#identity	#camera	#images	Resolution	Manual Annotation
VIPeR [7]	632	2	1.2k	48x128	Yes
3DPes	192	8	1k	37x88 – 236x278	Yes
ETH[8]	149	1	8.5k	44x85 – 175x449	Yes
iLIDS[9]	250	8	0.5k	32x76 – 115x294	Yes
CAVIARA[10]	72	2	1.2k	16x43 – 72x144	Yes
PRID[11]	200	2	4.4k	128x48	Yes
V47[12]	47	2	0.4k	51x154 – 222x446	Yes
WARD[13]	70	3	0.5k	48x128	No
CUHK02[14]	1816	10	7k	60x160	Yes
CUHK03[4]	1467	10	28k	31x92– 201x308	No
CUHK04[15]	8432	-	32k	19x53 – 141x375	Yes
RAiD[16]	43	4	7k	64x128	Yes
Shinpuhkan[17]	24	16	22k	48x128	Yes
MARS[18]	1261	6	59k	128x256	No
CMU[19]	33	31	13k	222x367 – 631x958	No
DUKE_MTMC [20]	1843	8	36k	64x156 –114x362	No
SAIVT[21]	150	8	7k	73x137 – 200x400	Yes

Table 2: Statistic of the people re-ID dataset. There are large variations in the image resolution, number of people, number of views for these datasets. For most large scale datasets (MARS, DUKE_MTMC), the groundtruth people bounding boxes were generated by computer algorithms, which heavily suffers from people misalignment.

Method	CUHK03	MARS	PRID	iLDS	ViPER	3dPES
Current arts	85.4[22]	77.4 [23]	43.6 [24]	64.6 [25]	**56.3** [22]	56.0 [25]
Our	**93.3**	**79.5**	**60.1**	**85.4**	52.5	**78.9**

Table 3: The top-1 retrieval accuracy of different methods. Except for ViPER, which is a small dataset with strong illumination and viewpoint variations, our method consistently outperforms the current arts by a margin.

Input	CUHK03	MARS	PRID	iLDS	ViPER	3dPES
RGB	91.1	76.9	55.0	84.5	42.4	70.0
RGB+KP	92.8	**79.8**	60.0	84.4	51.9	78.0
RGB+PAF	93.1	79.4	59.0	84.5	49.7	79.4
RGB+KP+PAF	**93.7**	**79.8**	**62.0**	**85.2**	**52.5**	**78.9**

Table 4: Ablative analysis of the pose heatmaps for the top-1 accuracy. Using all the heatmaps generated by CPM yields the best accuracy, albeit modest improvements over the key points (KP) or the part affinity fields (PAF) alone.

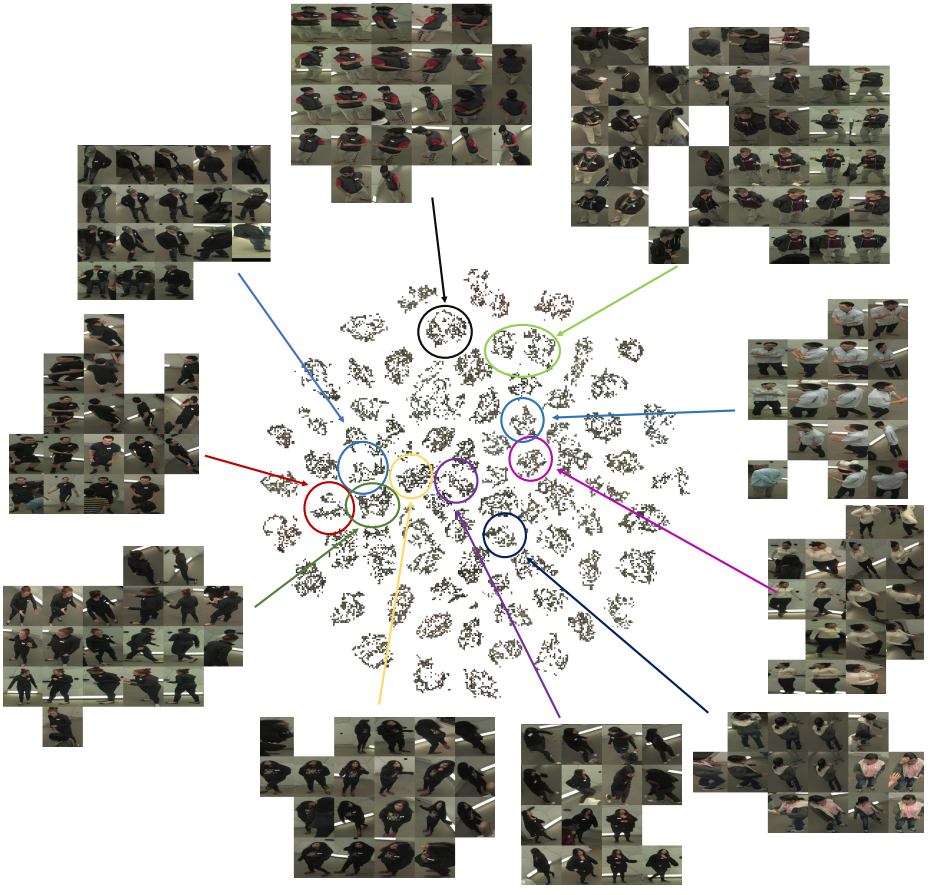


Fig. 6: The t-SNE visualization of our descriptor for 30k images of 80 people collected by the CMU Panoptic studio. Despite having many people with similar appearances, the images fo the same person are clustered together.

References

1. Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., Tang, X.: Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In: CVPR. (2017)
2. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR. (2017)
3. Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In: CVPR. (2016)
4. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: CVPR. (2014)
5. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR. (2015)
6. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML. (2015)
7. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. ECCV (2008)
8. Schwartz, W.R., Davis, L.S.: Learning discriminative appearance-based models using partial least squares. In: XXII Brazilian Symposium on Computer Graphics and Image Processing. (2009)
9. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: CVPR. (2010)
10. Cheng, D.S., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: BMVC. (2011)
11. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: Scandinavian conference on Image analysis. (2011)
12. Wang, S., Lewandowski, M., Annesley, J., Orwell, J.: Re-identification of pedestrians with variable occlusion and scale. In: ICCVW. (2011)
13. Martinel, N., Micheloni, C.: Re-identify people in wide area camera network. In: CVPRW. (2012)
14. Li, W., Wang, X.: Locally aligned feature transforms across views. In: CVPR. (2013)
15. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: CVPR. (2017)
16. Das, A., Chakraborty, A., Roy-Chowdhury, A.K.: Consistent re-identification in a camera network. In: ECCV. (2014)
17. Kawanishi, Y., Wu, Y., Mukunoki, M., Minoh, M.: Shinpuhkan2014: A multi-camera pedestrian dataset for tracking people across multiple cameras. In: 20th Korea-Japan Joint Workshop on Frontiers of Computer Vision. (2014)
18. Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., Tian, Q.: Mars: A video benchmark for large-scale person re-identification. In: ECCV. (2016)
19. Joo, H., Soo Park, H., Sheikh, Y.: Map visibility estimation for large-scale dynamic 3d reconstruction. In: CVPR. (2014)
20. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: ECCVW. (2016)
21. Bialkowski, A., Denman, S., Sridharan, S., Fookes, C., Lucey, P.: A database for person re-identification in multi-camera surveillance networks. In: International Conference on Digital Image Computing Techniques and Applications. (2012)

22. Geng, M., Wang, Y., Xiang, T., Tian, Y.: Deep transfer learning for person re-identification. arXiv preprint arXiv:1611.05244 (2016)
23. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
24. Paisitkriangkrai, S., Shen, C., van den Hengel, A.: Learning to rank in person re-identification with metric ensembles. In: CVPR. (2015)
25. Xiao, T., Li, H., Ouyang, W., Wang, X.: Learning deep feature representations with domain guided dropout for person re-identification. In: CVPR. (2016)
26. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. JMLR (2008)
27. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. TOG (2015)
28. http://mocap.cs.cmu.edu
29. Agarwal, S., Mierle, K., Others: Ceres solver. http://ceres-solver.org