Pose Switch-based Convolutional Neural Network for Clothing Analysis in Visual Surveillance Environment

Pendar Alirezazadeh [1,2], Ehsan Yaghoubi [2], Eduardo Assunção [2], João C. Neves [3], Hugo Proença [2]

Abstract: Recognizing pedestrian clothing types and styles in outdoor scenes and totally uncontrolled conditions is appealing to emerging applications such as security, intelligent customer profile analysis and computer-aided fashion design. Recognition of clothing categories from videos remains a challenge, mainly due to the poor data resolution and the data covariates that compromise the effectiveness of automated image analysis techniques (e.g., poses, shadows and partial occlusions). While state-of-the-art methods typically analyze clothing attributes without paying attention to variation of human poses, here we claim for the importance of a feature representation derived from human poses to improve classification rate. Estimating the pose of pedestrians is important to fed guided features into recognizing system. In this paper, we introduce pose switch-based convolutional neural network for recognizing the types of clothes of pedestrians, using data acquired in crowded urban environments. In particular, we compare the effectiveness attained when using CNNs without respect to *human poses* variant, and assess the improvements in performance attained by pose feature extraction. The observed results enable us to conclude that pose information can improve the performance of clothing recognition system. We focus on the key role of pose information in pedestrian clothing analysis, which can be employed as an interesting topic for further works.

Keywords: Soft biometrics, pedestrian clothing analysis, surveillance environment, human pose classification.

1 Introduction

The analysis of the pedestrian appearance, and more specifically clothing analysis, has gained interest in machine learning technologies in order to increase accuracy of surveillance based recognition systems. Clothing is one of the most important soft biometrics to pedestrian analysis and has many different applications, such as clothing retrieval [Li16a, Li16b], clothing recognition [IBK18, Li16c], outfit recommendation [TYO17] and visual search for matching fashion items [LBV18]. Despite several works proposed in clothing analysis, clothing recognition can't be considered a solved task, especially for surveillance-based environment, that typically produce poor quality data. A good clothing recognition system is highly dependent on the training phase. If these systems are trained with images in controlled conditions, they will not achieve high performance in the real world with various clothing appearance, styles and poses.

One of the major problems in the analysis of clothing is the lack of comprehensive dataset with enough images. Recently two datasets have been published. The MVC Dataset [LCC16]

[1] Pendar.Alirezazadeh@ubi.pt
[2] IT-Instituto de Telecomunicações, Portugal
[3] TomiWorld, Portugal

for view-invariant clothing retrieval with 161,638 images and the DeepFashion Dataset [Li16c] with 800,000 annotated real-life images. Both datasets are image-based dataset. Nowadays with cities getting bigger and increasing the use of city-level scenes, researchers have shown an increased interest in clothing analysis of pedestrians which are captured by cameras in streets [Hu19, YY11].

To perform clothing analysis in surveillance environment with uncontrolled conditions, we collected a dataset composed of video-based images from outdoor and indoor advertisement panels in Portugal and Brazil. On the other hand, clothing attribute analysis is highly dependent on deformation and poses variation of the human body. By moving some parts of the body such as the knee, hip, neck, shoulder etc. in various gestures, different types of clothing may look like each other, which causes the similarity of the extracted feature vectors and decreases the classification rate. In order to have the ability to clothing recognition in the real application, in this paper, we consider switching CNN architecture that passes frames from a video within a surveillance environment on related Pose-CNN based on a pose-switch classifier. The related Pose-CNN is chosen based on pose information extracted from the video frames as in multi-column Pose-CNN networks to augment the ability to confront pose variations. A particular Pose-CNN is trained on a video frame if the performance of the network on the frame's pose is the best. Fig. 1 illustrates the architecture of our proposed approach.

2 Pose identification

The pose identification aims to explore the human pose group, to assist convolutional neural network for better pedestrian clothing recognition. The output of pose identification is a pose number based on feature vector including a set of coordinates to describe the pose of the person. It consists of two main steps, including estimates human poses and classifies poses to select the appropriate network.

2.1 Human pose estimation

Human pose estimation also known as key-point detection, aims to detect the locations of K key-points or part of the body e.g. R-hip, L-hip, R-shoulder, L-shoulder etc. from bounding box images. So we have estimated K heatmaps where each heatmap indicates the location confidence of the defined key-point. In order to obtain pedestrian bounding boxes (BBs), we use the effective object detection technique VGG-based SSD 512 as pedestrian detector. Pedestrian BBs are fed into pose estimator and key points are generated automatically. In this paper we use CNN based Single Person Pose Estimator (SPPE) method to estimates poses. SPPE network is designed to train on single person images and it is very sensitive to localization errors [Fa17]. On the other hand, pose information consists of a set of key points that each key point belongs to specific region. To select region of interests which have high quality for SPPE network, we use Spatial Transformer Networks (STN) [Ja15]. The STN has shown excellent performance in modeling the variance of scale and

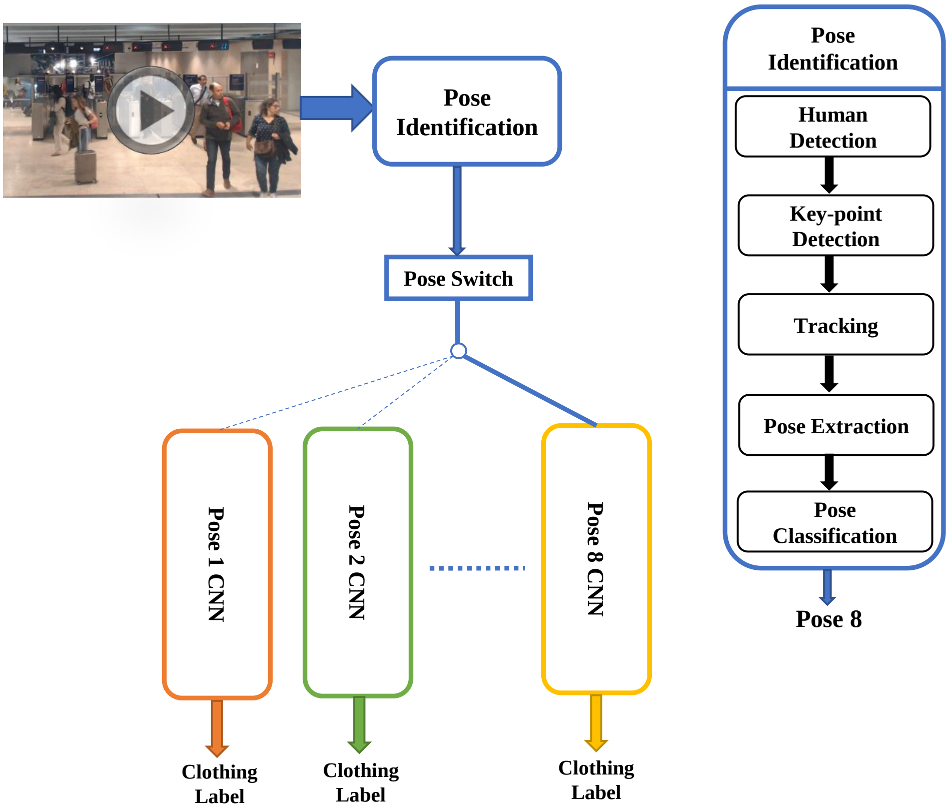

Fig. 1: Architecture of the proposed method, Pose Switch-CNN is shown. Video frames from the surveillance environment are relayed to one of the eight CNN networks based on the pose label inferred from pose identification.

pose for adaptively region localization [Li18]. The STN performs a 2D pointwise transformation with the affine parameters θ which can be expressed as:

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \tag{1}$$

where (x_i^t, y_i^t) are the target coordinates of the regular grid in the output feature map and the (x_i^s, y_i^s) are the source coordinates in the input feature map that define the sample points. The output of the SPPE network is a set of 16 key points which are used to pose estimation.

After human poses estimation for each BBs, we use pose similarity to track multi-person poses in videos to indicate the same person across different frames. Pose metric similarity is used to eliminate the poses which are too close and too similar to each others. We used intra-frame d_f and inter-frame d_c pose distance metrics to measure the pose similarity between two poses P_1 and P_2 in a frame and two sequential frames[Xi18]:

$$\begin{aligned} &d_f\left(P_1,P_2|\sigma_1,\sigma_2,\lambda\right) = K_{\mathrm{Sim}}\left(P_1,P_2|\sigma_1\right)^{-1} + \lambda H_{\mathrm{Sim}}\left(P_1,P_2|\sigma_2\right)^{-1}, \\ &d_c\left(P_1,P_2\right) = \textstyle\sum_{n=1}^{N} \frac{f_2^n}{f_1^n} \end{aligned} \tag{2}$$

where

$$K_{\mathrm{Sim}}\left(P_1,P_2|\sigma_1\right) = \begin{cases} \sum_{n=1}^{N} \tanh\frac{c_1^n}{\sigma_1} \cdot \tanh\frac{c_2^n}{\sigma_1} & \text{if } p_2^n \text{ is within } B\left(p_1^n\right) \\ 0;\ \mathit{otherwise} \end{cases} \tag{3}$$

$$H_{Sim}\left(P_1,P_2|\sigma_2\right) = \sum_{n=1}^{N} \exp\left[-\frac{\left(p_1^n - p_2^n\right)^2}{\sigma_2}\right] \tag{4}$$

where p_1^n and p_2^n are the n^{th} key points of pose P_1 and P_2 in $B\left(p_1^n\right)$ and $B\left(p_2^n\right)$ boxes respectively, N=16 is number of body keypoints, f_1^n and f_2^n are feature point extracted from boxes, and σ_1, σ_2 and λ can be determined in a data-driven manner. We have extracted coordinates(x,y) for 16 key-points of the full body for all the images. Then, these 16 coordinates points are concatenated to generate a 32 dimensional body coordinate-features vector for each human BBs.

2.2 Pose classification

Posed-based features may not necessarily be numerically similar for similar motions [Ya11] and it is an important challenge in pose-based feature applications. One of the practical solutions is finding a suitable pattern that aims to grouping a set of pose-based feature in such a way that features in the same group are more similar to each other than to those in other groups. For this purpose in this study we have used K-means classification algorithm. In order to raise the accuracy of the K-means, we use T-distributed Stochastic Neighbor Embedding (t-SNE) [ZWT18] method before classification. This method is known as a nonlinear dimensionality reduction technique for visualization high-dimensional data in a low-dimensional space of two or three dimensions that similar feature vectors are modeled by nearby points and dissimilar feature vectors are modeled by distant points with high probability. The t-SNE method aims to best capture neighborhood identity by considering the probability that one point is the neighbor of all other points. Conditional neighborhood probability of object x_i with object x_j is defined as:

$$p_{j|i} = \frac{\exp\left(-\left\|x_i - x_j\right\|^2/2\tau_i^2\right)}{\sum_{k\neq i}\exp\left(-\left\|x_i - x_k\right\|^2/2\tau_i^2\right)}, \tag{5}$$

where τ_i^2 is the variance for the Gaussian distribution centered around x_i. Since p_{ij} is not necessarily equal to p_{ji}, because τ_{ij} is not necessarily equal to τ_{ji}, so joint probabilities p_{ij} is defined by symmetrizing two conditional probabilities as:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}. \tag{6}$$

We have trained K-means with low dimension feature vectors resulted from t-SNE method and classified the body coordinate-features to K classes.

3 Results and Discussion

In this section, we briefly introduce the datasets, implementation details and results of the proposed method and comparison methods. The experimental results empirically validate the effectiveness of the proposed method.

3.1 Dataset

Due to the lack of comprehensive dataset for pedestrian clothing analysis in surveillance environment, we collected Biometria e Deteção de Incidentes (BIODI) dataset. The BIODI dataset collected from 216 videos recorded by 36 advertisement panels in Portugal and Brazil. These videos captured in various indoor and outdoor environments such as roads, beaches, airports, streets and metro stations at different hours of the day, lighting, pose, style and various weathers. In each panel, a camera is placed at a distance of 1.5 meters from the ground. All cameras have the same brand with different adjustments, which lead to videos with different qualities. There was no precondition and all of the videos were recorded in unconstraint environments. The statistics of BIODI dataset are summarized in the Table I. To recognize the enormous upper-body and lower-body clothing items, we have labeled the BBs manually. We generated category list bikini, blouse, coat, hoodie, shirt and t-shirt for the upper-body part and jean, legging, pant and short for the lower-body part. Each image received at most one category label for each part.

Factors	Statistics
No. of Videos	216
Length of Videos	7 minutes
Frame rate extraction	7 frames/sec.
No. of Subjects	13876
No. of Bounding Boxes (BBs)	503433
Aspect ratio of BBs (Height/Width)	1.75

Tab. 1: Statistics of the BIODI dataset

To further show the efficacy of our proposed methods, we conducted clothing recognition experiments on the RAP-2.0 [Li19] dataset and compared our results with the performance of their best method. RAP-2.0 comes from a realistic HighDefinition (1280 × 720) surveillance network at an indoor shopping mall and all images are captured by 25 cameras scenes. This dataset contains 84928 images (2589 subjects) with resolution ranging from 33 × 81 to 415 × 583.

3.2 Implementation Details

We have adopted K=8 typical poses, empirically. We consider a subset of 300,000 images of BIODI as training data and a subset of 100,000 images as validation data. Based on pose identification method, the training and validation data are divided to 8 typical pose groups. Clothes bounding boxes for upper-body and lower-body are detected by use of extracted key points. Time performance of pose identification algorithm for a frame including 20 people is about 0.3 second. To evaluate the performance of our proposed system after pose identification, we adopt end-to-end CNN approaches as clothing recognition. End-to-end deep learning methods have made jointly learn features and classifiers. We use CNNs with same architectures for each pose group. We fine-tune VGG-16 [SZ14] and ResNet50 [He16] on training and validation data with weights of ImageNet [Ru15] dataset for each pose group. In testing, we employ the remain part of BIODI to test the fine-tuned models. We ensure that no subject BBs overlaps between fine-tuning and testing sets. The stochastic gradient descent (SGD) is adopted to optimize the networks. For both models, we use the initial learning rate 1×10^{-4} and weight decay with 1×10^{-6}. The models have implemented in Python 3.6.7 using the Keras 2.1.6 deep learning library on top of the Tensorflow 1.10 backend and trained for 100 epochs with one NVIDIA GeForce RTX 2080 Ti GPU.

3.3 Results

We present an extensive evaluation of our proposed method on upper-body and lower-body clothing recognition. We firstly compare our framework with two baseline models (i.e. VGG-16 and ResNet50 without pose information) on BIODI dataset to validate the effectiveness of Pose Switch-based CNN. Table 2, 3 show the classification accuracy of baseline models on upper-body and lower-body BIODI clothing categories recognition, respectively. From the obtained results, the proposed technique increased the performance of clothing recognition rate on all pose groups of upper-body and lower-body parts. In order to visualize performance of the proposed framework which has had better performance compared to the situation without pose information, we have drawn the receiver operating characteristic (ROC) curve per category for the VGG-16 network (Fig.2). We have drawn the ROC curve for coat and blouse classes from upper-body and pants and short classes for lower-body. As it derives from the ROC curves, performance of VGG-16 network is improved using pose information. Secondly, to further show the efficacy of our approach, we conducted clothing recognition experiments on RAP-2.0 dataset and compared our results with the performance of the baseline method which is achieved best recognition rate. Based on the full body's direction, RAP-2.0 images are annotated to four types of viewpoints, including facing front (F), facing back (B), facing left (L) and facing right (R). Due to this background, we have classified images to four typical pose groups. The results of employing VGG-16 network in each of 4 typical pose groups on RAP-2.0 upper-body and lower-body parts are shown in Table 4, 5, respectively. It is clear that for all typical pose groups, the recognition rates are improved compared to without pose information, significantly.

Network	Pose 1	Pose 2	Pose 3	Pose 4	Pose 5	Pose 6	Pose 7	Pose 8	Mean Accuracy	without Pose
VGG-16	88.94	88.98	89.52	88.79	88.93	89.59	88.54	88.20	88.93	87.41
ResNet50	88.02	87.43	87.54	87.44	88.13	88.14	87.37	87.14	87.65	86.23

Tab. 2: The performance of the proposed method (%) for BIODI upper-body

Network	Pose 1	Pose 2	Pose 3	Pose 4	Pose 5	Pose 6	Pose 7	Pose 8	Mean Accuracy	without Pose
VGG-16	87.5	85.21	87.34	88.13	87.97	86.62	85.42	87.62	86.98	86.15
ResNet50	86.89	85.66	87.03	88.44	88.04	85.68	85.43	87.54	86.83	85.17

Tab. 3: The performance of the proposed method (%) for BIODI lower-body

Network	Pose 1	Pose 2	Pose 3	Pose 4	Mean
VGG-16	82.66	82.89	83.44	83.84	83.20
DeepMAR-R[Li19]	-	-	-	-	76.68

Tab. 4: The performance of the proposed method and deepMAR-R (%) for RAP-2.0 upper-body

4 Conclusion and Future Works

Since surveillance-based images are collected in unconstrained environment with various pose and styles, different types of clothing may look like each other which causes the similarity of the extracted feature vectors and decreases the classification rate. In this paper, we propose pose switch-based convolutional neural network that leverages pose variation to improve the accuracy of the pedestrian clothing recognition in crowded urban environments. The proposed method employs pose estimation techniques to key point detection for coordinate-features representation. We have classified all BBs to eight typical pose groups using these features. The convolutional neural networks are trained for each pose group and recognized upper-body and lower-body clothing images. Extensive experiments on RAP-2 datasets show that our method exhibits state-of-the art performance on major dataset in real surveillance scenarios. In the future, we plan to extend the proposed method to explore more efficient human semantic structure knowledge to assist pedestrian attribute recognition.

5 Acknowledgement

This research is funded by the “FEDER, Fundo de Coesão e Fundo Social Europeu” under the “PT2020 - Portugal 2020” program, “IT: Instituto de Telecomunicações” and “TOMI:

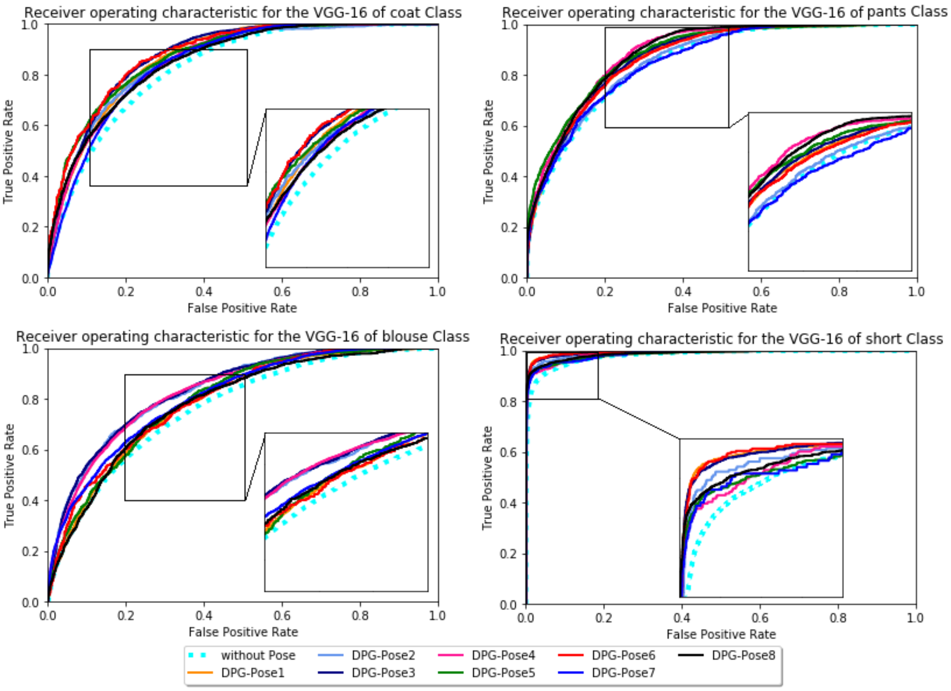

Fig. 2: ROC curves of the VGG-16 for different categories on BIODI upper-body and lower-body.

Network	Pose 1	Pose 2	Pose 3	Pose 4	Mean
VGG-16	87.13	86.93	87.49	87.06	87.15
DeepMAR-R[Li19]	-	-	-	-	81.33

Tab. 5: The performance of the proposed method and deepMAR-R (%) for RAP-2.0 lower-body

City's Best Friend". Also, the work is funded by FCT/MEC through national funds and when applicable co-funded by FEDER PT2020 partnership agreement under the project UID/EEA/50008/2019.

References

[Fa17] Fang, Hao-Shu; Xie, Shuqin; Tai, Yu-Wing; Lu, Cewu: Rmpe: Regional multi-person pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2334–2343, 2017.

[He16] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778, 2016.

[Hu19] Huang, Jin; Wu, Xinglong; Zhu, Jianlin; He, Ruhan: Real-Time Clothing Detection with Convolutional Neural Network. In: Recent Developments in Intelligent Computing, Communication and Devices, pp. 233–239. Springer, 2019.

[IBK18] Ivanov, Alexander Y; Borzunov, Georgii I; Kogos, Konstantin: Recognition and identification of the clothes in the photo or video using neural networks. In: 2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus). IEEE, pp. 1513–1516, 2018.

[Ja15] Jaderberg, Max; Simonyan, Karen; Zisserman, Andrew et al.: Spatial transformer networks. In: Advances in neural information processing systems. pp. 2017–2025, 2015.

[LBV18] Lasserre, Julia; Bracher, Christian; Vollgraf, Roland: Street2Fashion2Shop: Enabling Visual Search in Fashion e-Commerce Using Studio Images. In: International Conference on Pattern Recognition Applications and Methods. Springer, pp. 3–26, 2018.

[LCC16] Liu, Kuan-Hsien; Chen, Ting-Yen; Chen, Chu-Song: Mvc: A dataset for view-invariant clothing retrieval and attribute prediction. In: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval. ACM, pp. 313–316, 2016.

[Li16a] Li, Zongmin; Li, Yante; Tian, Weiwei; Pang, Yunping; Liu, Yujie: Cross-scenario clothing retrieval and fine-grained style recognition. In: 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, pp. 2912–2917, 2016.

[Li16b] Liang, Xiaodan; Lin, Liang; Yang, Wei; Luo, Ping; Huang, Junshi; Yan, Shuicheng: Clothes co-parsing via joint image segmentation and labeling with application to clothing retrieval. IEEE Transactions on Multimedia, 18(6):1175–1186, 2016.

[Li16c] Liu, Ziwei; Luo, Ping; Qiu, Shi; Wang, Xiaogang; Tang, Xiaoou: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1096–1104, 2016.

[Li18] Li, Dangwei; Chen, Xiaotang; Zhang, Zhang; Huang, Kaiqi: Pose Guided Deep Model for Pedestrian Attribute Recognition in Surveillance Scenarios. In: 2018 IEEE International Conference on Multimedia and Expo (ICME). IEEE, pp. 1–6, 2018.

[Li19] Li, Dangwei; Zhang, Zhang; Chen, Xiaotang; Huang, Kaiqi: A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. IEEE transactions on image processing, 28(4):1575–1590, 2019.

[Ru15] Russakovsky, Olga; Deng, Jia; Su, Hao; Krause, Jonathan; Satheesh, Sanjeev; Ma, Sean; Huang, Zhiheng; Karpathy, Andrej; Khosla, Aditya; Bernstein, Michael et al.: Imagenet large scale visual recognition challenge. International journal of computer vision, 115(3):211–252, 2015.

[SZ14] Simonyan, Karen; Zisserman, Andrew: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[TYO17] Tangseng, Pongsate; Yamaguchi, Kota; Okatani, Takayuki: Recommending outfits from personal closet. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2275–2279, 2017.

[Xi18] Xiu, Yuliang; Li, Jiefeng; Wang, Haoyu; Fang, Yinghong; Lu, Cewu: Pose flow: Efficient online pose tracking. arXiv preprint arXiv:1802.00977, 2018.

[Ya11] Yao, Angela; Gall, Juergen; Fanelli, Gabriele; Van Gool, Luc: Does human action recognition benefit from pose estimation? In: BMVC 2011-Proceedings of the British Machine Vision Conference 2011. 2011.

[YY11] Yang, Ming; Yu, Kai: Real-time clothing recognition in surveillance videos. In: 2011 18th IEEE International Conference on Image Processing. IEEE, pp. 2937–2940, 2011.

[ZWT18] Zhou, Hongyu; Wang, Feng; Tao, Peng: t-Distributed Stochastic Neighbor Embedding Method with the Least Information Loss for Macromolecular Simulations. Journal of chemical theory and computation, 14(11):5499–5510, 2018.