Image-based Anomaly Detection within Crowds

Thomas Golda

Vision and Fusion Laboratory
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT), Germany
thomas.golda@kit.edu

Technical Report IES-2018-02

Abstract

Authorities and security services have to deal with more and more data collected during events and on public places. Two reasons for that are the rising number of huge events, as well as the expanding coverage with CCTV cameras of areas within cities. Even the number of ground crew teams, that are equipped with mobile cameras, rises continuously. These examples show that modern surveillance and location monitoring systems come with need of suited assistance systems, which help the associated security workers to keep track of the situations. In this report, we present a first idea how such a system using modern machine learning algorithms could look like. Furthermore, a more detailed look on two state-of-the-art methods for human pose estimation is given. These algorithms are then investigated for their performance on the target domain of crowd surveillance scenarios using a small dataset called CrowdPose.

1 Introduction

In the first part of this report, the topic of image-based anomaly detection within crowds is motivated, followed by a short characterization of the target domain.

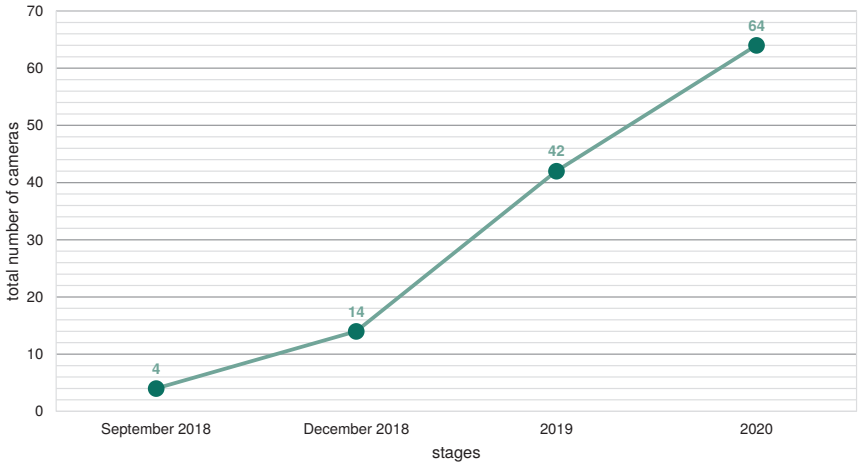

Figure 1.1: The number of cameras in the city of Mannheim will increase until 2020 up to 64 cameras and even more will follow.

1.1 Motivation

With the rising number of CCTV cameras equipped with high-resolution[1] image sensors, the task of monitoring public areas gets more and more difficult. On the one hand side, since most cameras are IP cameras, the data gathered produces large amounts of network traffic and storage utilization. On the other hand side, it is impossible for a single person to keep track of the situations within all connected cameras. As an example, the city and police of Mannheim, Germany decided to start a research project on the topic ”Intelligent Video Surveillance”. In cooperation with the *Fraunhofer-Institute for Optronics, System Technologies and Image Exploitation* the city center gets equipped with multiple cameras. The aim of the project in Mannheim is to do research on methods for analysing behavior of pedestrians and recognizing their activities. These methods should act as an assisting technology to help the security staff to do their job. The diagram displayed in Figure 1.1 shows how the number of cameras will develop within the

[1] 1280×720 and higher

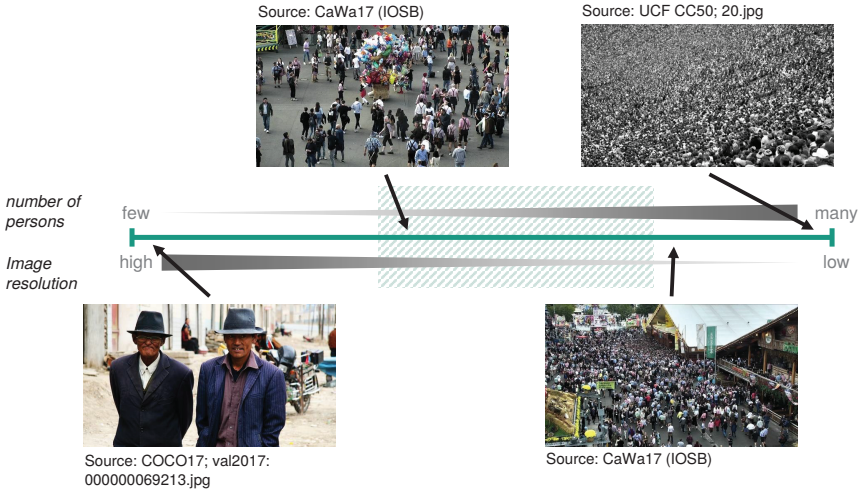

Figure 1.2: Pictures of people can range from portraits with just a single person, up to whole crowds, where the number of people shown is so high, that it is not possible anymore to detect single persons. The green hatched area shows the target domain on which this and future work will focus.

years 2018 to 2020. Another example is the Cannstatter Volksfest, an annual three-week beer festival in Stuttgart, Germany. In 2017 eleven surveillance cameras where used to monitor most of the area. The next year the number was increased by four cameras. Those examples show, that the number of cameras used at events and within cities is growing. This is a challenge not only for the security staff that has to keep an eye over all cameras, but also for the system itself, which has to cope with the large amount of data.

1.2 Characterization of the target domain

Surveillance cameras are used in various and heterogeneous environments. These are ranging from small shops, over hospitals and museums, up to large areas like the festival ground in Stuttgart where the Cannstatter Wasen takes place. Such scenarios differ in different ways, like e.g. in lighting conditions, privacy aspects and the size of the area monitored. A consequence of the last mentioned point, is the strongly varying size of single persons within the recorded video material.

Whereas a surveillance camera installed in a small shop typically records persons having a quite large size with recognizable facial features, a camera installed on a festival ground just records people with a size of only a few pixels. Figure 1.2 illustrates this situation. Pictures and video material showing people can range from single portrait pictures to images of dense crowds. Due to the strongly varying size of a single person and the different amounts of dynamic occlusions, developing a suited method is a challenging task. Therefore, it is necessary to reduce the complexity of the initial domain, which is done by concentrating on the typical views of static surveillance cameras, where some dozens of people are present.

2 Related work

Anomaly detection is an important topic in various fields, like the analysis of continuous and discrete time series [SGPE17], surveillance video streams [SCS18] and medical imaging [TCSOM+09]. All approaches have in common that they aim to develop a representation of some kind of default situation, which then is compared to the current one in order to decide whether it is an abnormal or normal situation. This is mostly done using machine learning algorithms, especially unsupervised and semi-supervised ones. The methods in the field of anomaly detection within the context of surveillance scenarios can be divided into three main categories: reconstruction models, predictive modeling, and deep generative models. [KTP18] These will be presented shortly in the following.

2.1 Reconstruction Models

This category of methods uses some intermediate representation generated from the original data. Linear and non-linear methods like principal component analysis (PCA) and Autoencoders (AE) are used to generate these representations from appearance or motion, which model the normal behavior in surveillance videos. Some representatives from this category are [XRY+15], which uses stacked de-noising Autoencoders (SDAE) to generate a representation based on input image and optical flow, [HCN+16], which uses Spatio-Temporal SDAEs on multiple stacked frames to generate a representation, and [VPN+17], which use Deep Belief Networks (DBN) for the generation of a representation. The latest

publications are dominated by SDAEs, since they allow localization of anomalies compared to classical PCA and AE. [KTP18]

2.2 Predictive Models

In contrast to reconstruction models, which have the goal to learn a generative model that can reconstruct frames of a video, the goal of predictive models is to predict the current frame as a function of its predecessors. Some methods that can be counted to this category are LSTM-based methods like [WLG17] and [TBWW17], combining AEs and LSTMs, or [WS02], which use Slow Feature Analysis (SFA) that aims to extract slowly varying representations of rapidly varying high dimensional input. [KTP18]

2.3 Deep Generative Models

The last category mainly consists of Variational Autoencoders (VAE) [AC15], Generative Adversarial Networks (GAN) [DVR+18] and adversarially trained AutoEncoders (AAE), which are used for the purpose of modeling the likelihood of normal video samples in an end-to-end deep learning framework. Especially in the context of image- and video-based anomaly detection, GANs are used. The basic idea in anomaly detection is to be able to evaluate the density function of the normal vectors in the training set containing no anomalies while for the test set a negative loglikelihood score is evaluated, which serves as the final anomaly score. The score corresponds to the test sample's posterior probability of being generated from the same generative model representing the training data points. GANs provide a generative model that minimizes the distance between the training data distribution and the generative model samples without explicitly defining a parametric function, which is why it is called an implicit generative model. [KTP18]

3 Human Pose Estimation for Anomaly Detection

In Section 2 we gave an overview over existing work on anomaly detection. However, many methods like [DVR+18] and [RDFS11] use global motion context like dense optical flow for the analysis of video sequences. This is done

to detect anomalies implicitly, since it is hard to tell what an anomaly looks like, beforehand. The first draft of our approach is displayed in Figure 3.1. Starting with an input image or sequence of images, the workflow consists of three major parts: the estimation of human body poses, the extraction of motion information based on (sparse) optical flow methods, and in the end a classification of the motion information. In this report, we focus on the first part shown in the schematics, namely the estimation of human body poses. The remaining parts will be part of future work.

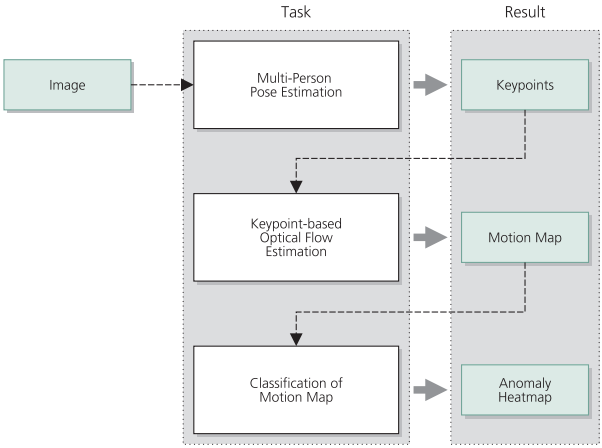

Figure 3.1: Illustration of the conceptual idea for anomaly detection based on human pose estimation. First, pose estimation is performed on the input material. The obtained key points are then tracked and classified in order to detect anomalies.

3.1 OpenPose

OpenPose [CHS+18] is a framework for multi-person pose estimation. It is based on the method presented in [CSWS17], which follows a multi-stage approach. First, it generates so called *Confidence Maps* for the estimation of body key points. These maps contain information about the distribution of particular key point types within the input image. For each type of key point one Confidence Map is computed, containing estimated locations for all key points of this type within the image. Second, so called *Part Affinity Fields* (PAF) are generated. These are used

for the corresponding connections between key points and contain the information, which key points might belong together, based on visible limbs or the human body. The information obtained by computing PAFs is then used in a third and last step to cluster and connect key points belonging to the same person. This is done using the Hungarian Method [KY55]. Since the algorithm's first step consists of trying to detect all key points within an image, it can be counted as a *bottom-up* method. The major benefits of bottom-down methods are their computational speed and scalability with regard to the number of persons.

3.2 AlphaPose

AlphaPose is another framework for human pose estimation. Different from OpenPose, the underlaying method belongs to the group of top-down methods. This method was presented in [FXTL17] and its main idea is to detect humans and perform single-person pose estimation on each detections, which follows the typical workflow of top-down methods. The detection-driven approach is also the main benefit of top-down methods, since they perform better for single and small persons. The main problem tackled by [FXTL17] is the avoidance of multiple pose proposals for a single person, caused by several detections of the same person. In order to achieve this, the proposed solution uses non-maximum suppression to choose the best suggested pose and iteratively removes similar pose estimates. Furthermore, as presented in [XLW+18], taking time into account improves the performance, since poses are connected over consecutive time steps. In combination with non-maximum suppression over time, this leads to more robust poses.

4 Experiments

In order to investigate how modern algorithms for image-based human pose estimation perform on data taken from the target domain, a small dataset was created. The results obtained by our experiments on this dataset and the dataset itself are presented in the following.

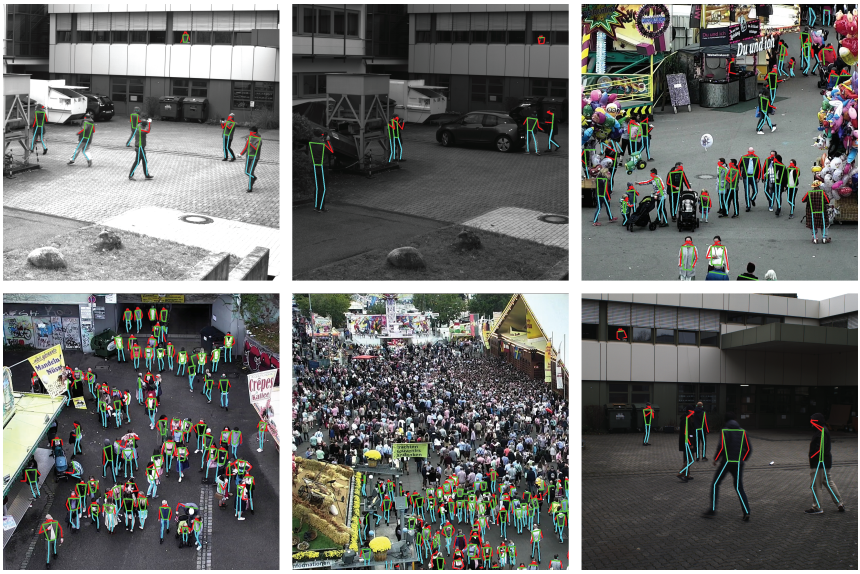

Figure 4.1: Examples showing patches from images taken from the CrowdPose [Dis18] dataset with corresponding annotations. The dataset consists of pictures showing different illumination situations, person sizes, number of persons and viewing angles.

4.1 CrowdPose Dataset

CrowdPose [Dis18] is a small dataset consisting of 25 different images with each image having a size of about two megapixels. In total, 833 persons were annotated over all pictures. Each image is annotated with at least six and at most 148 individuals. Figure 4.1 shows some exemplary patches taken from pictures of the CrowdPose dataset. For the annotation of the collected images we used the open source tool *sloth*.[2] We therefore developed an extension for sloth, which allowed us to annotate images with key points, automatically generated appropriate bounding boxes, person ids and activities. Despite the fact that CrowdPose is much smaller compared to other existing datasets labeled for human pose estimation like the COCO [LMB+14] and MPII [APGS14] dataset, it also differs significantly

[2] `https://github.com/cvhciKIT/sloth`

in the number of persons per image. Whereas COCO has an average number of about two persons and a maximum number of 13 persons per image, CrowdPose comes with about 33 and 148 respectively. [Ron17]

4.2 Quantitative evaluation

For the quantitative evaluation we decided to adapt the *Object Keypoint Similarity*[3] (OKS) used for the evaluation on the COCO dataset. Equation (4.1) shows the formula for the OKS.

$$\text{OKS} = \frac{\sum_i \exp(\frac{-d_i^2}{2s^2\kappa_i^2})\delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \tag{4.1}$$

The d_i are the Euclidean distances between corresponding detections and ground truth points and the $v_i \in \{0, 1, 2\}$ are the visibility flags given by the dataset in order to disregard occluded keypoints from the metric. This is controlled by δ which is defined as $\min\{v_i, 2\}$. The remaining two variables are the object scale s, which is defined as the square root of the segmentation area, and a keypoint constant κ_i that has been determined by computing the standard deviation of humans by annotating multiple images redundantly. [Ron17]
Since CrowdPose does not provide any information about semantic segmentation, we could not use OKS as proposed. In order to solve this problem, we adapted OKS and replaced the segmentation area in s by the area of the resulting tightly fitted bounding boxes provided by CrowdPose. Table 4.1 shows the impact of our adapted OKS on the derived average precision ($\overline{\text{AP}}$) and average recall ($\overline{\text{AR}}$). [Dis18]

[3] `http://cocodataset.org/#keypoints-eval`

Table 4.1: The adapted metrics $\overline{\text{AP}}$ and $\overline{\text{AR}}$ show a similar values compared to AP and AR that were determined using the original OKS. [Dis18]

	OpenPose	**AlphaPose**
$\text{AP}_{@OKS=0.50:0.95}$	0.29912	0.41202
$\overline{\text{AP}}_{@OKS=0.50:0.95}$	0.38154	0.46619
$\text{AR}_{@OKS=0.50:0.95}$	0.32849	0.42581
$\overline{\text{AR}}_{@OKS=0.50:0.95}$	0.40806	0.47581

Since the $\overline{\text{AP}}$ and $\overline{\text{AR}}$ are in competing range to AP and AR, we were encouraged to use the adapted OKS for our experiments.
Table 4.2 shows evaluation results on CrowdPose for OpenPose and AlphaPose. AlphaPose beats OpenPose in all experiments achieving up to 8.3 times higher performance. It is conspicuous, that all obtained results are much lower compared to the evaluation results on COCO dataset displayed in Table 4.1.

Table 4.2: The table shows $\overline{\text{AP}}$ / $\overline{\text{AR}}$ for AlphaPose and OpenPose on CrowdPose dataset. Both methods were evaluated on the whole CrowdPose dataset, as well as on both its subsets. AlphaPose outperforms OpenPose in all experiments. [Dis18]

	Combined	**Cannstatter Wasen**	**IOSB**
AlphaPose	0.00349 / 0.01297	0.00006 / 0.00359	0.01249 / 0.06103
OpenPose	0.00088 / 0.00324	0.00003 / 0.00043	0.00231 / 0.01765

The main reason for this can be found in the evaluation process itself. As presented earlier, OKS uses some key point constants, which were obtained using annotated images from COCO dataset. If we compare the appearance of CrowdPose images and those from COCO, we can see that those from CrowdPose differ significantly to those from the latter. It seems that the determined key point constants cannot be used directly for the evaluation on CrowdPose. Especially the adaption using the scale s might have a strong influence on the sensitivity of the metric to the size of a single person and hence the actual key point locations. For all experiments, the methods were used without any changes and without further fine-tuning on the target domain.

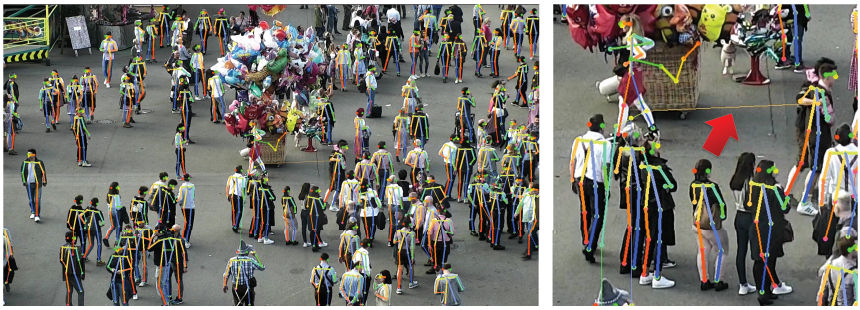

Figure 4.2: Exemplary result generated with AlphaPose. On the the first glance, the left side shows promising results. The right part shows a central patch taken from the left image. The red arrow indicates a wrong connection between key points of different persons, which is just one of multiple wrong poses within the patch.

4.3 Qualitative evaluation

Despite the strong discrepancy of the quantitative evaluation results between the evaluation on COCO and CrowdPose presented in Section 4.2, the qualitative results show promising results. Figure 4.2 shows an evaluation example generated using AlphaPose. At first glance, the result looks good. However, when we take a more detailed look on this example, we can see some apparent failures. The most salient one is indicated by the red arrow: obviously, a wrong connection has been predicted between two persons. Furthermore, especially when two or more persons overlap, often body skeletons are predicted over multiple persons. Two examples within the patch are the two guys on the right hand side, and the two girls in the bottom right part. Nonetheless, the evaluation shows that human pose estimation is suited for the application on the target domain. In order to improve the results, the state-of-the-art methods have to be slightly adapted in order to be more robust against overlapping persons and obviously wrong inter-connections between unrelated persons. These problems will be tackled in future work.

5 Summary and Future Work

In this report, we presented an overview over existing methods for image-based anomaly detection within crowded scenarios. Furthermore, we introduced the field of human pose estimation and evaluated two state-of-the-art algorithms for their performance on the CrowdPose dataset, which was created to investigate the methods apart from typical application scenarios. The two algorithms, OpenPose [CHS+18] and AlphaPose [FXTL17], representing the two main approaches for human pose estimation performed similarly well. The broadly used metric OKS used by the COCO keypoint challenge [LMB+14], reports quite bad results on the own dataset. However, the qualitative evaluation showed promising results. In future work we will mainly concentrate on three different essential aspects that came up during our first experiments: the adaption of existing methods for human pose estimation to the target domain of crowded scenarios, a mathematical definition of anomalies in crowded scenarios and the application of human pose estimation algorithms for anomaly detection.

Bibliography

[AC15] Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. 2015.

[APGS14] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d Human Pose Estimation: New Benchmark and State of the Art Analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[CHS+18] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018.

[CSWS17] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[Dis18] Thomas Dissert. *Posenerkennung von Personen innerhalb von Menschenmengen*. KIT, Karlsruhe, 2018.

[DVR+18] Lucas Deecke, Robert Vandermeulen, Lukas Ruff, Stephan Mandt, and Marius Kloft. Anomaly detection with generative adversarial networks, 2018.

[FXTL17] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017.

[HCN+16] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis. Learning temporal regularity in video sequences. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 733–742, 2016.

[KTP18] B. Ravi Kiran, Dilip Mathew Thomas, and Ranjith Parakkal. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *Journal of Imaging*, 4(2), 2018.

[KY55] H. W. Kuhn and Bryn Yaw. The hungarian method for the assignment problem. *Naval Res. Logist. Quart*, pages 83–97, 1955.

[LMB+14] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[RDFS11] D. Ryan, S. Denman, C. Fookes, and S. Sridharan. Textures of optical flow for real-time anomaly detection in crowds. In *2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 230–235, Aug 2011.

[Ron17] Matteo Ruggero Ronchi. COCO 2017 Keypoints Challenge, October 2017.

[SCS18] W. Sultani, C. Chen, and M. Shah. Real-world anomaly detection in surveillance videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, June 2018.

[SGPE17] Dominique T. Shipmon, Jason M. Gurevitch, Paolo M. Piselli, and Stephen T. Edwards. Time series anomaly detection; detection of anomalous drops with limited features and sparse examples in noisy highly periodic data. *CoRR*, abs/1708.03665, 2017.

[TBWW17] Eleni Tsironi, Pablo Barros, Cornelius Weber, and Stefan Wermter. An analysis of convolutional long short-term memory recurrent neural networks for gesture recognition. *Neurocomput.*, 268(C):76–86, December 2017.

[TCSOM+09] Alberto Taboada-Crispi, Hichem Sahli, Maykel Orozco Monteagudo, Denis Hernández-Pacheco, and Alexander Falcon. *Anomaly Detection in Medical Image Analysis*, pages 426–446. 01 2009.

[VPN+17] Hung Vu, Dinh Q. Phung, Tu Dinh Nguyen, Anthony Trevors, and Svetha Venkatesh. Energy-based models for video anomaly detection. *CoRR*, abs/1708.05211, 2017.

[WLG17] W. Liu W. Luo and S. Gao. Remembering history with convolutional lstm for anomaly detection. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, 2017.

[WS02] Laurenz Wiskott and Terrence J. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Comput.*, 14(4):715–770, April 2002.

[XLW+18] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose Flow: Efficient online pose tracking. In *BMVC*, 2018.

[XRY+15] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. Learning deep representations of appearance and motion for anomalous event detection. *CoRR*, abs/1510.01553, 2015.