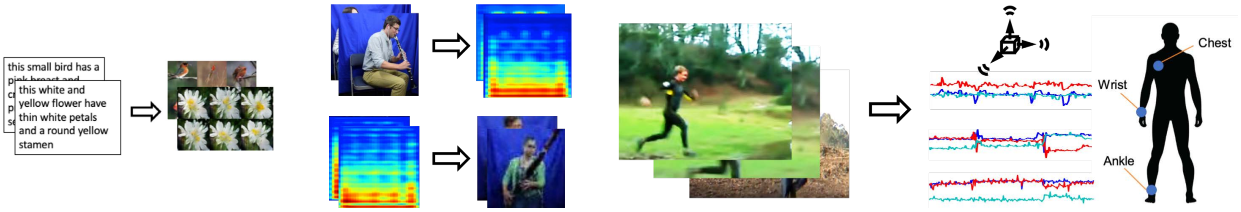
Deep Generative Cross-modal On-body Accelerometer Data Synthesis from Videos

Shibo Zhang
Northwestern University
Department of Computer Science
shibo.zhang@northwestern.edu

Nabil Alshurafa
Northwestern University
Department of Preventive Medicine and Department of Computer Science
nabil@northwestern.edu


this small bird has a pink breast and ...
this white and yellow flower have thin white petals and a round yellow stamen
Chest
Wrist
Ankle

(a) Text to image

(b) Video to audio and audio to video

(c) Video to accelerometer data

Figure 1: Multi-modal generation tasks with different modality pairs. (a) Convert text descriptions into images [20]. (b) Generate audio from videos and generate videos from audio [5]. (c) The proposed accelerometer data generation from videos.

ABSTRACT

Human activity recognition (HAR) based on wearable sensors has brought tremendous benefit to several industries ranging from healthcare to entertainment. However, to build reliable machine-learned models from wearables, labeled on-body sensor datasets obtained from real-world settings are needed. It is often prohibitively expensive to obtain large-scale, labeled on-body sensor datasets from real-world deployments. The lack of labeled datasets is a major obstacle in the wearable sensor-based activity recognition community. To overcome this problem, I aim to develop two deep generative cross-modal architectures to synthesize accelerometer data streams from video data streams. In the proposed approach, a conditional generative adversarial network (cGAN) is first used to generate sensor data conditioned on video data. Then, a conditional variational autoencoder (cVAE)-cGAN is proposed to further improve representation of the data. The effectiveness and efficacy of the proposed methods will be evaluated through two popular applications in HAR: eating recognition and physical activity recognition. Extensive experiments will be conducted on public sensor-based activity recognition datasets by building models with synthetic data and comparing the models against those trained from real sensor data. This work aims to expand labeled on-body sensor data, by generating synthetic on-body sensor data from video, which will equip the community with methods to transfer labels from video to on-body sensors.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing design and evaluation methods**; • **Computing methodologies** → **Unsupervised learning**; **Neural networks**; **Artificial intelligence**.

KEYWORDS

Deep Multi-modal Learning; Video-sensor Data Representation Learning; Deep Generative Model; Accelerometer Data Synthesis; Data Augmentation

ACM Reference Format:
Shibo Zhang and Nabil Alshurafa. 2020. Deep Generative Cross-modal On-body Accelerometer Data Synthesis from Videos. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers (UbiComp/ISWC '20 Adjunct), September 12–16, 2020, Virtual Event, Mexico.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3410530.3414329

1 PROBLEM STATEMENT

With the prevalence of wearable devices in our daily life, human activity recognition (HAR) based on wearables has emerged as a novel approach for fitness tracking and wellness monitoring. However, the lack of labeled large-scale datasets is a major obstacle in the sensor-based activity recognition community. For example, for physical activity recognition, most commonly used public datasets only contain labeled sensor data ranging from 15 hours (in 29 subjects) [12] to 30 hours (in 10 subjects) [19], and it has been shown that increasing the amount of data used to train the

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

UbiComp/ISWC '20 Adjunct, September 12–16, 2020, Virtual Event, Mexico

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-8076-8/20/09…$15.00
https://doi.org/10.1145/3410530.3414329

model (including generated augmentation data) leads to further improved results [17, 23]. Nowadays, deep learning has exhibited extraordinary discriminative and generative power for modeling complex data in a plethora of application domains, such as image recognition, video understanding, speech recognition, and machine translation among other domains [8]. In the era of deep neural networks, without large-scale labeled datasets, it is difficult to leverage the benefit of deep learning. Therefore, exploring methods to acquire labeled datasets efficiently has been a research interest among many research communities for a long period of time. For sensor-based activity recognition, the situation is even more challenging given the fact that, unlike images or audio, from which the annotation can be obtained from the raw data, annotation of sensor data is difficult for humans to do without the use of video recordings post-experimentally. Ground truth acquisition requires burdensome self-report, the presence of observers, or the use of video recording that captures the activities of participants [1]. Nevertheless, currently existing methods require large human effort to obtain labeled datasets. Therefore, an efficient and effective method to quickly acquire large-scale, fully labeled datasets would be useful to the research community.

Currently, in addition to spending more resources collecting larger scale datasets, data augmentation methods have been studied to achieve better performance on HAR tasks [13, 17, 23]. Most current data augmentation solutions only contain artificially warping, scaling or jittering the real sensor data; thus the heterogeneity of generated data is limited [17, 23]. Kwon et al. [13] recently utilized data of a different modality (i.e., video) to generate on-body sensor data in a sophisticated engineered pipeline. However, the method suffers from robustness issues such as vigorous movement, change of scenery, and occlusion. Moreover, the requirement for heavy adjustment limits its wide deployment in real-world tasks. Nevertheless, the idea of employing data of a highly related modality to synthesize motion sensor data is genuinely brilliant, in that machine learning communities with interest in different modalities can share knowledge, as well as datasets. The idea of cross-modal transferring inspired the work presented here.

To overcome the problems mentioned above, a possible solution is to use multi-modal representation learning to generate synthetic data from learnt multi-modal distributions, which has proven effective in other modalities [26]. Multi-modal representation learning can enable a number of applications across a variety of modalities, including image captaining [10], image generation from text [20], and conversion from video to audio and vice versa [5], as shown in Figure 1. However, multi-modal representation learning with video and IMU sensor data has only been sparsely studied [16]. To my knowledge, there has not yet been any work on deep multi-modal generative networks targeting sensor data synthesis, especially accelerometer data synthesis based on video data. In this work, I aim to fill the gap between deep multi-modal generative models and sensor data synthesis.

This work presents two deep generative cross-modal models for on-body accelerometer data synthesis and demonstrates the usability and efficacy via extensive experiments. I will conduct experiments in two tasks: eating recognition and physical activity recognition. The models will be trained using video and on-body accelerometer sensor, and tested on public sensor-based activity recognition datasets. The usability of the generated sensor data will be validated by comparing the activity recognition performance of models trained with synthetic data and with real data.

2 RELATED WORK

Work closest to mine falls under two categories: sensor data synthesis and deep multi-modal generative model, relatively in perspective of problem-wise and algorithm-wise.

Sensor Data Synthesis: Besides the long standing problem for general time series data augmentation, recently there has been growing interest in sensor data augmentation techniques [2, 13, 17, 23, 24]. Some of them augment the sensor data by artificially warping, adding noise to, or impose spatial transformation to the real sensor data [17, 23]. Thus, the augmented data, which are in essence distorted or transformed original data, has limited heterogeneity with the 'seed' data. Therefore, these methods have limited capability in generating sensor data of multi-factor heterogeneity. Alzantot et al. [2] employed GAN to synthesize sensor data using existing sensor data. However the quality of synthesized data was not quantitatively evaluated but merely judged by loss function as well as another real/synthetic classifier. In another work, Wang et al. [24] also utilized GAN for sensor data synthesis using sensor data. The quantitative performance didn't show convincing results of their proposed method. Kwon et al. [13] employed a highly hand engineered pipeline to extract the 3D motion from video to generate virtual on-body sensor data. The complex pipeline requires careful adjustment and parameter fine-tuning. In contrast, I propose end-to-end models to learn the multi-modal representation from the raw data and generate synthetic sensor data.

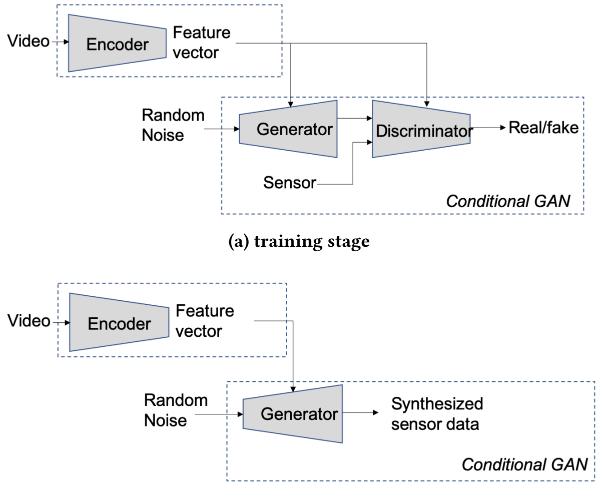
Deep Multi-modal Generative Model: Although few researchers use deep generative model to build a video-sensor multi-modal system, however, if we take a broader view, we can find a list of works that use multi-modal generative model in other pairs of modalities such as audio and video [5], text and image [10, 20]. To keep the conciseness of this paper, we review the most relevant and typical works in recent years. Chen et al., proposed a deep cross-modal audio-visual generation model to generate audio data from video and generate video data from audio using URMP dataset [5]. Yang et al. [25] used a temporal model to jointly build a deep multi-modal model and model temporal sequential information at the same time. What's worth mentioning, Nakamura et al. used a stacked LSTM, which takes multi-modal video and acceleration features as input and builds a multi-task discriminative model for simultaneous activity recognition and energy expenditure estimation [16].

3 PRELIMINARIES

This section briefly introduces two deep learning architectures employed in this work: Variational Autoencoder (VAE) and Generative Adversarial Network (GAN).

3.1 Variational Autoencoder (VAE)

VAE [11] is a widely used deep generative model in representation learning. VAE builds the mapping from the input observation x to a compressed code z_s in the manner of an encoder, and then decodes the coding to reconstruct the observation. The latent representation



(a) training stage

(b) test stage

Figure 2: The architecture of Algorithm 1 in (a) training stage and (b) test stage.

is calculated through:

$$z_s = \mu_{\boldsymbol{x}} + \sigma_{\boldsymbol{x}} * \boldsymbol{\epsilon} \tag{1}$$

with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 1)$. VAE imposes the code $\boldsymbol{z_s}$ on a Gaussian distribution:

$$\bar{p}(\boldsymbol{z_s}) = \mathcal{N}(z_s|\mathbf{0}, \boldsymbol{I}) \tag{2}$$

The latent representation $\boldsymbol{z_s}$ is supposed to learn the representative attributes of the input raw data, thus to be useful for data generation.

3.2 Generative Adversarial Network (GAN)

GAN [9] is composed of two components - generator (F_G) and discriminator (F_D). The Generator (F_G) and Discriminator (F_D) are competing with each other as a zero-sum game framework, in the manner that F_G aims at confusing the discriminator and F_D tries to distinguish the samples generated by F_G and the samples from the original dataset. Both F_G and F_D are competing to individually become more powerful of imitating original data samples and discrimination capability iteratively. Thus the distribution of the data is learned by the generator. A standard GAN has no control over the modes of the data to generate [15]. The objective function is

$$\min_G \max_D V(D, G) = \mathbb{E}_{x\sim p_{data}(x)}[logD(x)] + \mathbb{E}_{x\sim p_z(z)}[log(1 - D(G(z)))] \tag{3}$$

where $p(data)$ is the target data distribution and z is drawn from a random noise distribution $p(z)$; $G(z)$ is the sample produced by the generator; $D(x)$ is the probability emitted by discriminator that x is a real example rather than a fake one drawn from the model.

4 METHODOLOGY

In this section, I present two proposed approaches that utilize deep cross-modal generative models in accelerometer data synthesis.

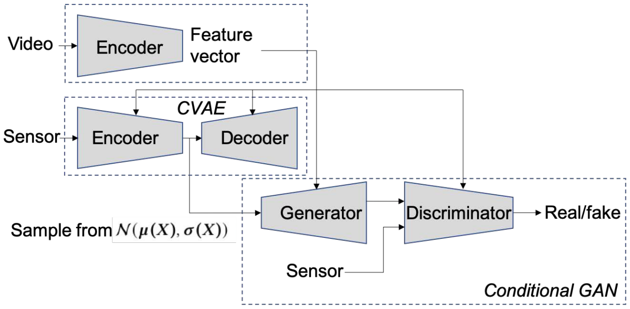


Figure 3: The architecture of Algorithm 2

4.1 Algorithm Design 1

The first design uses a conditional GAN to build a generative cross-modal model. The architecture consists of two parts: *video encoder* and *conditional GAN* (cGAN).

A video encoder is used to condense the high-dimensional video clip into a feature vector. The *I3D* model [3] has powerful discriminative capability and flexible input configuration for both raw video and optical flow data, making it state-of-the-art model for visual activity understanding in many sub-tasks. Therefore *I3D* model is a candidate for video encoder. Besides *I3D* model, other video activity recognition models as well as human pose estimation models will also be tested.

A cGAN [15] models the conditional distribution of sensor data based on video representation. cGAN takes extra information as additional input on which the learned data distribution is conditioned. Multiple types of information can be used to be conditioned on, such as class label, information from another modality, or even itself. Similar to Equation 3, the objective function of cGAN is

$$\min_G \max_D V(D, G) = \mathbb{E}_{x\sim p_{data}(x)}[logD(x|y)] + \mathbb{E}_{x\sim p_z(z)}[log(1 - D(G(z|y)))] \tag{4}$$

where $p(data)$ is the target distribution and $p(z)$ is a random distribution; $G(z|y)$ is the sample produced by the generator conditioned on y; $D(x|y)$ is the probability that x is a real sample given by discriminator conditioned on y.

Here the video feature vector from video encoder is taken as additional information to be conditioned on. The intuition is that the accelerometer signal depends on the movement of the subject in the video, which can be extracted and encoded in a feature representation. Specifically, the generator and discriminator of GAN can be in the form of LSTM to leverage its sequential data modeling capability. After the model is trained, in test stage, the video encoder takes the video as input and yields the visual representation. The generator of cGAN takes as input the random Gaussian noise $\mathcal{N}(0, 1)$, and uses the visual representation to condition on. The output of the generator is the synthesized sensor data.

4.2 Algorithm Design 2

In this section, conditional VAE-conditional GAN (cVAE-cGAN) is presented. Different from Algorithm 1, a sensory cVAE is added to generate the distribution from which the random noise input of cGAN is sampled. The video encoder and cGAN are the same as

Algorithm 1. A conditional Variational Autoencoder (cVAE) utilizes the information from video to learn a conditional distribution of the sensor data. The idea stems from the intuition that a meaningful prior distribution, which is the latent representation generated by cVAE, will improve the generative capability of GAN. The structural choice of encoder and decoder in cVAE could be fully connected network [28].

The test stage is similar to that in Algorithm 1, only except that the generator of cGAN takes as input the random noise sampling from the sensor data distribution learned by VAE.

5 EVALUATION

The proposed methods will be evaluated on two tasks: eating recognition and physical activity recognition. There are three types of dataset in this study: *training dataset* that includes both labeled video and sensor data, *test dataset* containing videos with activity annotation, and *validation dataset* which is sensor-based activity dataset. Two deep generative models will be built for each task on the training set, and a large number of accelerometer data will be generated from videos in the test dataset. Note that the synthesized sensor data shares the same annotation with the video, thus sensor data is inherently annotated. The synthesized sensor data will be utilized in the training of sensor-based activity classifiers, and trained classifier will be tested on validation dataset. Several widely used classifiers including Logistic Regression (LR), Random Forest (RF), Adaboost, Convolutional Neural Network (CNN), as well as activity recognition model DeepConvLSTM [18] will be chosen. These activity classifiers will be trained using either only synthesized accelerometer data, only real accelerometer data or synthesized data combined with a small part of real data, in order to validate the effectiveness of the synthesized accelerometer data.

5.1 Task 1: Eating Recognition

This task aims at exploring the data generation capability of the proposed model through a fine-grained classification problem, while synthesizing missing sensor data in two self-collected datasets.

To ensure the training dataset has a high degree of diversity in terms of the activity type, activity scene, as well as heterogeneity of accelerometer sensor, three dataset are combined as training set: CMU-MMAC Dataset [6], our iSenseOvereating eating dataset [27] and smoking&eating dataset, all of which contain third-party videos and wrist-worn accelerometer data. CMU-MMAC Dataset contains both egocentric videos and third-party videos; accelerometer sensor placement includes arms, legs and back; wrist-worn accelerometer data are also collected. In this task only third-party videos and wrist-worn accelerometer will be used. Our iSenseOvereating and smoking&eating dataset have totally around 70 subjects and 50 hours third-party video in lab, both with fine-grained annotation for eating and smoking gestures. Part of the wrist-worn IMU sensor data are missing due to hardware failure. The part with sensor data will be used in training dataset, and the part without sensor data as test dataset. After the model is trained, the synthesized data will be validated on two eating detection datasets [14, 22] for feeding gesture recognition task in the aforementioned scheme.

5.2 Task 2: Physical Activity Recognition

The goal is to synthesize chest-worn sensor data to assist developing a model to recognize physical activity (standing, walking, running, jumping jack, etc) from wrist-worn sensor stream. Three datasets are used as training set: Stanford-ECM Dataset [16], CMU-MMAC Dataset [6], and the Sense2StopSync dataset [29]. Stanford-ECM Dataset comprises about 27 hours egocentric video and chest-mounted accelerometer data. For CMU-MMAC Dataset, the egocentric videos and accelerometer at the back will be used here. The Sense2StopSync dataset contains 45.2 hours of egocentric video from 21 participants in free-living conditions, and the participant sensor suite includes a chest-worn acceleormeter. Test dataset includes ThirdToFirst Dataset [7], which has egocentric videos with physical activity annotation, and YouTube egocentric videos, of which the titles and descriptions can be served as annotation. The synthesized data will be tested using popular sensor-based activity recognition datasets including PAMAP2 [21] and Opportunity [4].

Further, ablation study for both tasks will be conducted on VAE-cGAN and cVAE-GAN which separately has the conditional distribution modeling removed from cVAE and cGAN, in order to evaluate the effect of the conditional inference of cVAE-cGAN.

6 EXPECTED CONTRIBUTION

Overall, the design and implementation of two video-based deep generative on-body accelerometer data synthesis models will be presented. In this work, a list of contributions to the community can be achieved:

- Two end-to-end deep cross-modal accelerometer data generation models based on video data are proposed, which can be used to produce synthesized sensor data given video containing human activities. This work aims at addressing the large scale labeled IMU sensor data scarcity problem by utilizing existing labeled video datasets.
- The proposed deep generative models will be evaluated through experiments on multiple datasets. The proposed models performance will be evaluated through comparing the accuracy of activity recognition classifiers trained on purely synthesized sensor data, real sensor data, and mixed synthesized and real sensor data.
- A sensor data latent representation space will be learned conditioned on video data, which can enable other video-senor based multi-modal tasks, such as activity detection, recognition and time synchronization.

The learned representation will also endow and motivate other application scenarios such as cross-modal data retrieval, semi-supervised learning, etc. I believe a larger group of applications related to activity analysis will be unlocked based on the knowledge and findings in this work.

7 BIOGRAPHICAL SKETCH

Shibo Zhang is a PhD student in the HABitsLab at Northwestern University, supervised by Prof. Nabil Alshurafa. His research interest is human activity recognition and machine learning. He received an MS in computer science from Northwestern University, Evanston, IL, USA, in 2017, where he is currently pursuing a PhD in computer science.

REFERENCES

[1] Rawan Alharbi, Tammy Stump, Nilofar Vafaie, Angela Pfammatter, Bonnie Spring, and Nabil Alshurafa. 2018. I Can't Be Myself: Effects of Wearable Cameras on the Capture of Authentic Behavior in the Wild. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 90 (Sept. 2018), 40 pages. https://doi.org/10.1145/3264900

[2] Moustafa Alzantot, Supriyo Chakraborty, and Mani Srivastava. 2017. SenseGen: A deep learning architecture for synthetic sensor data generation. *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)* (Mar 2017). https://doi.org/10.1109/percomw.2017.7917555

[3] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 6299–6308.

[4] Ricardo Chavarriaga, Hesam Sagha, Alberto Calatroni, Sundara Tejaswi Digumarti, Gerhard Tröster, José del R. Millán, and Daniel Roggen. 2013. The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters* 34, 15 (2013), 2033 – 2042. https://doi.org/10.1016/j.patrec.2012.12.014 Smart Approaches for Human Action Recognition.

[5] Lele Chen, Sudhanshu Srivastava, Zhiyao Duan, and Chenliang Xu. 2017. Deep Cross-Modal Audio-Visual Generation. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017* (Mountain View, California, USA) *(Thematic Workshops '17)*. Association for Computing Machinery, New York, NY, USA, 349–357. https://doi.org/10.1145/3126686.3126723

[6] Fernando de la Torre, Jessica K. Hodgins, Javier Montano, and Sergio Valcarcel. 2009. Detailed Human Data Acquisition of Kitchen Activities: the CMU-Multimodal Activity Database (CMU-MMAC). In *CHI 2009 Workshop. Developing Shared Home Behavior Datasets to Advance HCI and Ubiquitous Computing Research.*

[7] Mohamed Elfeki, Krishna Regmi, Shervin Ardeshir, and Ali Borji. 2018. From Third Person to First Person: Dataset and Baselines for Synthesis and Retrieval. arXiv:1812.00104 [cs.CV]

[8] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning.* MIT press.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27.* Curran Associates, Inc., 2672–2680.

[10] MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A Comprehensive Survey of Deep Learning for Image Captioning. *ACM Comput. Surv.* 51, 6, Article 118 (Feb. 2019), 36 pages. https://doi.org/10.1145/3295748

[11] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[12] Jennifer R. Kwapisz, Gary M. Weiss, and Samuel A. Moore. 2011. Activity Recognition Using Cell Phone Accelerometers. *SIGKDD Explor. Newsl.* 12, 2 (March 2011), 74–82. https://doi.org/10.1145/1964897.1964918

[13] Hyeokhyen Kwon, Catherine Tong, Harish Haresamudram, Yan Gao, Gregory D. Abowd, Nicholas D. Lane, and Thomas Ploetz. 2020. IMUTube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition. arXiv:2006.05675 [cs.CV]

[14] Konstantinos Kyritsis, Christina Lefkothea Tatli, Christos Diou, and Anastasios Delopoulos. 2017. Automated analysis of in meal eating behavior using a commercial wristband IMU sensor. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).* IEEE, 2843–2846.

[15] Mehdi Mirza and Simon Osindero. 2014. Conditional Generative Adversarial Nets. *CoRR* abs/1411.1784 (2014). arXiv:1411.1784 http://arxiv.org/abs/1411.1784

[16] Katsuyuki Nakamura, Serena Yeung, Alexandre Alahi, and Li Fei-Fei. 2017. Jointly Learning Energy Expenditures and Activities Using Egocentric Multimodal Signals. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 6817–6826.

[17] Hiroki Ohashi, M Al-Nasser, Sheraz Ahmed, Takayuki Akiyama, Takuto Sato, Phong Nguyen, Katsuyuki Nakamura, and Andreas Dengel. 2017. Augmenting wearable sensor data with physical constraint for DNN-based human-action recognition. In *ICML 2017 Times Series Workshop.* 6–11.

[18] Francisco Ordóñez and Daniel Roggen. 2016. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors* 16, 1 (Jan. 2016), 115. https://doi.org/10.3390/s16010115

[19] Daniele Ravì, Charence Wong, Benny P. L. Lo, and Guang-Zhong Yang. 2016. Deep learning for human activity recognition: A resource efficient implementation on low-power devices. *2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN)* (2016), 71–76.

[20] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative Adversarial Text to Image Synthesis. In *Proceedings of The 33rd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 48)*, Maria Florina Balcan and Kilian Q. Weinberger (Eds.). PMLR, New York, New York, USA, 1060–1069. http://proceedings.mlr.press/v48/reed16.html

[21] Attila Reiss and Didier Stricker. 2012. Introducing a New Benchmarked Dataset for Activity Monitoring. In *Proceedings of the 2012 16th Annual International Symposium on Wearable Computers (ISWC) (ISWC '12).* IEEE Computer Society, USA, 108–109. https://doi.org/10.1109/ISWC.2012.13

[22] Edison Thomaz, Irfan Essa, and Gregory D Abowd. 2015. A practical approach for recognizing eating moments with wrist-mounted inertial sensing. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing.* 1029–1040.

[23] Terry T. Um, Franz M. J. Pfister, Daniel Pichler, Satoshi Endo, Muriel Lang, Sandra Hirche, Urban Fietzek, and Dana Kulić. 2017. Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. *Proceedings of the 19th ACM International Conference on Multimodal Interaction - ICMI 2017* (2017). https://doi.org/10.1145/3136755.3136817

[24] Jiwei Wang, Yiqiang Chen, Yang Gu, Yunlong Xiao, and Haonan Pan. 2018. SensoryGANs: An Effective Generative Adversarial Framework for Sensor-based Human Activity Recognition. In *2018 International Joint Conference on Neural Networks, IJCNN 2018, Rio de Janeiro, Brazil, July 8-13, 2018.* IEEE, 1–8. https://doi.org/10.1109/IJCNN.2018.8489106

[25] Xitong Yang, Palghat Ramesh, Radha Chitta, Sriganesh Madhvanath, Edgar A. Bernal, and Jiebo Luo. 2017. Deep Multimodal Representation Learning from Temporal Data. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Jul 2017). https://doi.org/10.1109/cvpr.2017.538

[26] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. 2020. Multimodal Intelligence: Representation Learning, Information Fusion, and Applications. *IEEE Journal of Selected Topics in Signal Processing* 14, 3 (Mar 2020), 478–493. https://doi.org/10.1109/jstsp.2020.2987728

[27] Shibo Zhang, William Stogin, and Nabil Alshurafa. 2018. I sense overeating: Motif-based machine learning framework to detect overeating using wrist-worn sensing. *Information Fusion* 41 (2018), 37 – 47. https://doi.org/10.1016/j.inffus.2017.08.003

[28] Xiang Zhang, Lina Yao, and Feng Yuan. 2019. Adversarial Variational Embedding for Robust Semi-supervised Learning. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining* (Jul 2019). https://doi.org/10.1145/3292500.3330966

[29] Yun Zhang, Shibo Zhang, Miao Liu, Elyse Daly, Battalio Samuel, Santosh Kumar, Bonnie Spring, James M. Rehg, and Nabil Alshurafa. 2020. SyncWISE: Window Induced Shift Estimation for Synchronization of Video and Accelerometry from Wearable Sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3, Article 107 (Sept. 2020), 27 pages. https://doi.org/10.1145/3411824