Multimodal Gait Recognition Under Missing Modalities

Ruben Delgado-Escano, Francisco M Castro, Nicolas Guil, Vicky Kalogeiton, Manuel J Marin-Jimenez

MULTIMODAL GAIT RECOGNITION UNDER MISSING MODALITIES

Rubén Delgado-Escaño
University of Malaga

Francisco M. Castro
University of Malaga

*Nicolás Guil**
University of Malaga

Vicky Kalogeiton
LIX, École Polytechnique,
CNRS, IP Paris

Manuel J. Marín-Jiménez
University of Cordoba,
IMIBIC

ABSTRACT

Multimodal systems for gait recognition have gained a lot of attention. However, there is a clear gap in the study of missing modalities, which represents real-life scenarios where sensors fail or data get corrupted. Here, we investigate how to handle missing modalities for gait recognition. We propose a single and flexible framework that uses a variable number of input modalities. For each modality, it consists of a branch and a binary unit indicating whether the modality is available; these are gated and merged together. Finally, it generates a single and compact 'multimodal' gait signature that encodes biometric information of the input. Our framework outperforms the state of the art on TUM-GAID and extensive experiments reveal its effectiveness for handling missing modalities even in the multiview setup of CASIA-B. The code is available online: `https://github.com/avagait/gaitmiss`.

1. INTRODUCTION

Like fingerprints or handwritten signatures, *gait* is a biometric feature that allows for people identification. Its main advantages are that it does not require the collaboration of the subject and can be performed at certain distance. Therefore, great effort has been put in its advancement [1, 2, 3]. Typical approaches use a single modality or input data type [4, 5, 6]. However, this is a big limitation, as nowadays it is easy to find devices (e.g. Kinect, mobiles) or techniques that produce different kinds of data like depth [7] or optical flow [8]. Thus, some works exploit multimodality and show that it leads to better representations and improved results [2, 9, 10, 11, 12]. Nevertheless, their common limitation is their inability to handle missing modalities. Specifically, they require all modalities at the same time and, therefore, they cannot be used in cases where one or more modalities are missing, such as sensor failures or data corruption; hence, they cannot be applied to several online real-life scenarios.

In contrast, we propose a single framework for gait recognition using a variable number of modalities (Fig. 1). It handles and combines various input modalities and is robust to *missing* ones at test time. Specifically, at training, it can fuse gray, optical flow and depth with one branch per modality (red, blue, green in Fig. 1), whereas at test time it can deal with sequences from a single RGB camera (no depth). This is essential for devices with computational or consumption constrains (such as cellular phones), where complex real-time computation is not feasible. In addition to the input data, each branch takes as input a binary value indicating if the input is available at test time (ellipses in Fig 1). The output of each branch is its corresponding signature. Finally, our framework combines the available signatures to produce a multimodal one, used to predict the identities.

Our contributions can be summarized as: *(i)* a novel multimodal gait recognition framework robust to missing modalities; *(ii)* generation of robust view-independent gait signatures; and, *(iii)* state-of-the-art results for TUM-GAID.

2. RELATED WORK

Gait recognition has been an active research topic for the past decades due to various applications, e.g. surveillance, crime prevention, forensic identification and social security.

Multimodal gait recognition. Most works use a single modality [1, 5, 13] based on silhouettes [1, 4], skeletons [14] or models [15, 16]. Common multimodal works rely on silhouettes [17, 18], optical flow [19], infrared [12], pose [2, 20] or depth [10]. Their common limitation is that they require all modalities both at training and test time, thus rending systems incapable of functioning when a modality is missing.

Missing modalities. Several recent works address the missing modalities issue by reconstructing the missing ones by using either AutoEncoders [21, 22] or GANs [23], while others [24] just focus on the problem of learning with missing modalities. In those cases, the sources are generally images/videos and text, applied to diverse problems. However, none of them targets gait recognition, where generating the missing gait input (a biometric source) is not an option.

*Thanks to the Junta de Andalucía (P18-FR-3130), the Ministry of Education (PID2019-105396RB-I00) of Spain, and to NVIDIA (GPU donation).

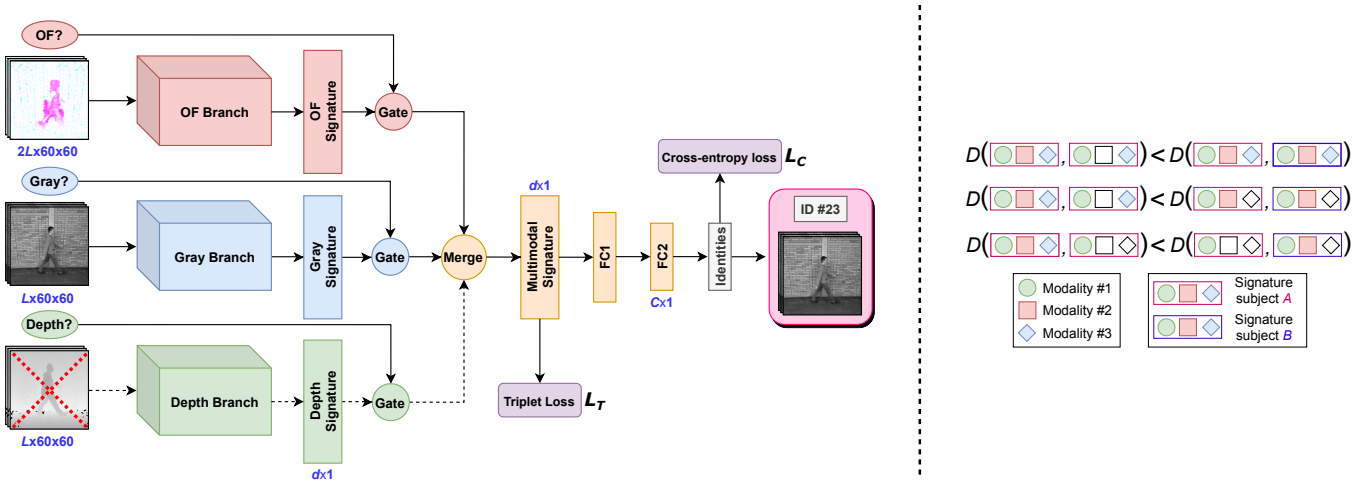

Fig. 1: **Multimodal gait recognition network robust to missing modalities. (Left)** Input: (*OF?, Gray?, Depth? ellipses*) binary input units indicating whether the modality is available – here, depth is not available (dashed red cross); (*volumes*) sequences of L frames for the different modalities. After fusing the single-modality signatures, a multimodal gait signature of d dimensions is further compressed by FC1. The final FC2 contains C classes (used just for training). **(Right)** At training, the network learns multimodal signatures so that the distance D between a pair of signatures of the same subject is lower than the distance between signatures of different subjects, independently of the modalities used to generate the signatures. To imitate test situations, some modalities are disabled (i.e. missing) at training (empty shapes).

3. PROPOSED MODEL

We propose a multimodal framework for gait recognition robust to missing modalities at test time (Fig. 1). It learns multimodal gait signatures that are similar for the same subject, and different for different subjects, regardless the combination of input data. Although several works combine multiple modalities [10, 2] and report improved results as the modalities offer complementary information [2, 11], this is not always realistic as in real life some modalities might be missing. However, to the best of our knowledge, no existing approach tackles missing modalities for gait recognition, i.e. *how to design a single model resilient to a variable number of available modalities*.

Overview. Our framework takes as input videos in various modalities and outputs the predicted identity of the person in the video (Fig. 1). It consists of three branches, one per input modality trained in parallel (optical flow, gray, and depth, Sec. 3.1). At test time, not all input modalities are required. To achieve this flexibility, we provide an additional input per branch indicating whether the modality is available or not (ellipses in Fig. 1). The output of each branch is passed through a *gate mechanism*, and these gated intermediate representations are merged to a single multimodal signature (*merge operation*). These two mechanisms together with the additional input allow the network to deal with missing modalities (Sec. 3.2). Finally, the multimodal signature is fed to the classification layers that predict the identities. Note that layer 'FC2' (Fig. 1) is used just for training, while 'FC1' is the gait signature directly used by any classifier (e.g. kNN).

3.1. Input Branches

Our framework consists of up to three branches, one per input modality: optical flow, gray, and depth. Note that it can be easily extended to any number of modalities.

Optical flow branch. We use an architecture similar to the "temporal stream" of [25], i.e. 2D conv and pooling layers, ending with fully-connected (FC). It is composed of four conv layers (96:7x7; 192:5x5; 512:3x3; 512:2x2) with ReLU and max-pooling (2x2) followed by two FC layers (2048; 1024).

Gray and depth branches. We define a fully-convolutional branch based on 3D convolutions to capture local temporal information. In particular, it consists of seven 3D conv layers (64:3x5x5; 128:3x3x3; 256:3x3x3; 512:3x3x3; 512:3x2x2; 512:2x1x1; 1024:1x1x1) with ReLUs in all layers but the last.

3.2. Gate mechanism and merge operation

Here, we describe the gate and merge operations that allow the network to deal with missing modalities.

Gate mechanism. Our framework takes as additional inputs k binary units u indicating whether one input modality is available or not. Each binary unit u_i acts as a **gate**, allowing or not the information coming from its corresponding modality to flow within the network.

Notations. For simplicity, we assume two input modalities m_1 and m_2; but extending to any number of inputs is straightforward. Let $\mathcal{B}_1$ and $\mathcal{B}_2$ be the backbone networks that extract features from m_1 and m_2, respectively. Let $\mathbf{f}_1$ and $\mathbf{f}_2$ of dimensionality d be the output vectors obtained from those backbones, respectively.

Merge operation. It takes as input the outputs of the gate mechanism. It is defined as an aggregation function $\rho(\cdot)$:

$$\rho(\mathbf{f}_1, u_1, \mathbf{f}_2, u_2, ..., \mathbf{f}_k, u_k) = \max(u_1 \cdot \mathbf{f}_1, u_2 \cdot \mathbf{f}_2, ..., u_k \cdot \mathbf{f}_k), \quad (1)$$

resulting in a d-dimensional output multimodal vector $\mathbf{z}$. The intuition behind this choice for $\rho(\cdot)$ is the following. When u_j is 1, the components of $\mathbf{f}_j$ will *compete* to be part of the output multimodal vector $\mathbf{z}$. In contrast, when u_j is 0, regardless the value of $\mathbf{f}_j$ that information will not become part of the output $\mathbf{z}$, and the modalities will have to *collaborate* to produce multimodal signatures that are similar even though one or more modalities are missing. In summary, a gait signature $\mathbf{z}$ from two modalities is obtained as:

$$\mathbf{z} = \rho\left(\mathcal{B}_1(m_1, \theta_1), u_1, \mathcal{B}_2(m_2, \theta_2), u_2\right) = \max(u_1 \cdot \mathcal{B}_1(m_1, \theta_1), u_2 \cdot \mathcal{B}_2(m_2, \theta_2)), \quad (2)$$

where θ_1 and θ_2 are the parameters of the $\mathcal{B}_1$ and $\mathcal{B}_2$ networks, respectively, that are learned during training. The vector $\mathbf{z}$ is L2-normalized before further processing.

3.3. Loss function for training

To deal with missing modalities, we use the Triplet Loss [26]: $\mathcal{L}_{\mathrm{T}}(A, P, N) = \max\left(\|g(A) - g(P)\|^2 - \|g(A) - g(N)\|^2 + \alpha, 0\right)$, where $g(\cdot)$ is a deep neural network, A is an anchor sample, P is a positive sample w.r.t. A, N is a negative sample w.r.t. A, and α is a margin value to be cross-validated. This ensures that gait signatures for samples of the same subject are similar (i.e. minimum distance), whereas the ones from different subjects are different (i.e. maximize their distance) regardless of the viewpoint, clothing or available modalities. Additionally, we use the Cross-Entropy loss $\mathcal{L}_{\mathrm{C}}$ to aid training the feature extractor. Finally, the overall loss $\mathcal{L}_{\mathrm{M}}$ is: $\mathcal{L}_{\mathrm{M}} = \mathcal{L}_{\mathrm{T}} + \beta \cdot \mathcal{L}_{\mathrm{C}}$, where β is a positive weight, experimentally chosen.

4. EXPERIMENTS

Here, we present the results of our framework. We use up to three input modalities with complementary information [11]: optical flow, gray and depth (when available). Unlike other methods that fine-tune their models on the gallery of the test partition [11, 14], we experiment *without* fine-tuning. We directly apply the pre-trained model on the test samples and classify them using a simple *k*NN classifier, thus validating the generalization of the feature extractor.

Pre-processing of input modalities. First, we spatially detect the subjects using Faster-RCNN [27], pre-trained on MS-COCO, and group them into tracks based on their feature distance. Then, following [5], we align the stacked subsequences of 25 maps so that the body is x-located in the middle of the central frame (i.e. #13) according to the obtained tracks. For better generalization, we extract samples from the video sequences with an overlap of 80% with the previous ones. Finally, we scale down the input maps to 80×60 pixels, keeping the original aspect ratio, and we remove any unnecessary background by cropping the maps to 60×60 (the full height is kept). The optical flow is obtained with SpyNet [8] pre-trained on MPI Sintel. The depth maps are represented as gray-scale images, i.e. scaling depth values to $[0, 255]$.

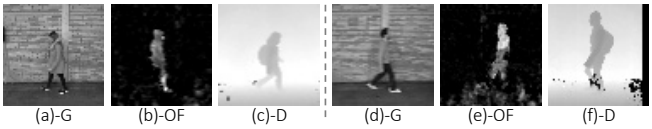

Fig. 2: Gait recognition results on three subjects from TUM-GAID using different modalities (a,d) gray, (b,e) optical flow, (c,f) depth. (d,e,f) samples are correctly predicted by using only one modality.

Implementation details. We implement our model using the Keras version of TensorFlow. We use cross-validation and set (α,β)= (1.0,0.1) in ($\mathcal{L}_{\mathrm{T}}$,$\mathcal{L}_{\mathrm{M}}$). The learning rate starts at 0.001 and is reduced by 0.2 when the validation loss plateaus. At training, each minibatch contains balanced samples of different covariate factors (e.g. normal, bag, shoes) and, for each sample, we have different versions of it, i.e. with all modalities, first modality missing, second modality missing, etc. Data augmentation is applied. To obtain a gait descriptor at video level, we average the descriptors obtained from samples of 25 frames (i.e. subsequences).

4.1. Datasets and metrics

We experiment with two datasets that provide RGB videos:

TUM-GAID [9]: It contains 305 subjects performing two walking trajectories indoors. Four situations are captured by a Microsoft Kinect: normal walk (*N*), carrying a backpack (*B*), wearing coating shoes (*S*) and, there is an elapsed time case where 32 subjects were recorded wearing different clothes (*TN-TB-TS*). We follow the train and test splits from [9]: 150 subjects for training and 155 subjects for testing.

CASIA-B [28]: It consists of 124 subjects that walk indoors. Actions are captured from 11 viewpoints (from $0°$ to $180°$ in steps of $18°$) with a resolution of 320×240 pix. Three situations are considered: normal walk (*NM*), wearing a coat (*CL*), and carrying a bag (*BG*). Following [3], we use the first 74 subjects at train and val and the last 50 at test; also, the target camera is not included in the gallery: identical-view cases are excluded for evaluating robustness to changes in viewpoint.

Metrics: We use Rank-1 (R1) accuracy, i.e. the percentage of correctly classified videos: $R1 = \#\text{correct}/\#\text{total}$.

4.2. Ablation study

We report in Tab. 2 the results obtained on individual samples of TUM-GAID when: (*i*) CE: we use only Cross-Entropy Loss in $\mathcal{L}_{\mathrm{M}}$ (no triplet); and (*ii*) SUM-Merge: the merge function is changed to the 'SUM' operator, i.e. sum of the gated descriptors. We observe that the use of the Triplet Loss is important. Then, MAX-merge brings also a small improvement.

Input Size	*Method*	*N*	*B*	*S*	*TN*	*TB*	*TS*	*Avg*
640 × 480	SiameseAE [14]	98.7	93.6	98.0	81.4	76.2	78.1	95.1
	PFM [29]	99.7	99.0	99.0	78.1	62.0	54.9	96.0
60 × 60	MTaskCNN [30]	99.7	97.4	99.7	59.4	62.5	68.8	95.6
	3D-CNN+Fusion [11]	100	99.4	99.4	75	62.5	62.5	96.5
	Ours (G+D & G+D+OF)	99.7	98.1	98.1	100	100	100	**98.8**

Table 1: **State of the art on TUM-GAID**. Rank-1 identification rate (%) at video level. Each column corresponds to a different scenario. Row 'G+D & G+D+OF' indicates that the same result is obtained with either all modalities or OF missing (G+D). Our results are obtained with a 3NN classifier on 256D signature.

Cases	*N*	*B*	*S*	*Avg*
CE	57.0	63.0	59.0	59.7
SUM-Merge	97.4	92.4	95.4	95.1
Ours: Full	97.8	93.0	96.0	**95.6**

Table 2: **Ablation study on TUM-GAID**. Rank-1 identification rate (%) at subsequence level. CE: Only cross-entropy loss without triplet loss in $\mathcal{L}_{\mathrm{M}}$ (Sec. 3.3); SUM-Merge: merge function is changed to the 'sum' operator. Columns: different scenarios.

Classifier-Modality†	*N*	*B*	*S*	*Avg*
(a) **BL-single**-G	98.9	95.1	96.0	96.7
(b) **BL-single**-OF	70.0	51.2	57.4	59.5
(c) **BL-single**-D	83.8	73.0	80.4	79.1
(d) **BL-all**-G+OF+D	95.8	91.4	91.2	92.8
(e) **BL-late**-G+OF+D	98.7	97.0	97.0	97.6
(f) **Ours**-G+OF+D	97.8	93.0	96.0	**95.6**

†G: Gray, OF: Optical Flow, D: Depth

Table 3: **Baselines on TUM-GAID: no missing.** Rank-1 identification rate (%) at subsequence level. Each row represents a different baseline approach.

Modalities: G: Gray, OF: Optical Flow, D: Depth						
G	D	OF	*N*	*B*	*S*	*Avg*
–	–	✓	78.5	64.8	71.8	71.7
–	✓	–	87.5	77.5	80.9	82.0
✓	–	–	94.3	87.1	90.7	90.7
–	✓	✓	87.6	77.7	80.9	82.1
✓	–	✓	94.4	87.2	90.7	90.8
✓	✓	–	97.6	92.6	95.6	95.3
✓	✓	✓	97.8	93.0	96.0	**95.6**

Table 4: **Missing modalities at test time on TUM-GAID**: Rank-1 identification rate (%) at subsequence level with 3NN on 256D signatures.

Modalities†					
G	OF	*NM*	*BG*	*CL*	*Avg*
–	✓	89.4	60.7	39.9	63.3
✓	–	99.1	86.1	36.1	73.8
✓	✓	99.6	89.5	45.5	**78.2**

†G: Gray, OF: Optical Flow

Table 5: **Missing modalities at test time on CASIA-B**: Rank-1 identification rate (%) at video level with 3NN on 2048D gait signatures. Note, CASIA-B contains only gray and optical flow (no depth).

4.3. Baseline models

We compare our framework to five baselines on TUM-GAID and report the results in Tab. 3. (a)-(c) BL-single: we train one model per modality with the same backbone and test with 3NN on 256D signatures (FC1 in Fig. 1); (d) BL-all: multimodal model with the same architecture as our model, where no modality is missing neither at train nor at test time, tested with 3NN on 256D signatures (FC1); (e) BL-late: three-branch multimodal model with no missing modality, where the modalities are combined with late fusion (with input the average of the softmax of each modality). We observe that gray is the most discriminative cue (a), as gray alone obtains high performance that cannot be beaten by either other modalities (b and c) or early fusion (d). However, our approach (f) improves BL-allas it uses samples with missing modalities during training thus leading to better generalization. Finally, the multimodal upper-bound BL-late (e) achieves the best performance but it requires all modalities present at test time.

4.4. Missing modalities

Our goal is to address the missing modalities issue. To reveal the effectiveness of our model, we experiment on two datasets: TUM-GAID and CASIA-B (multiview). For TUM-GAID, we report results at 'sample level' (i.e. 25-frame input) in Tab. 4. Gray brings the highest gain (third, seventh rows) whereas OF the smallest (first, fourth rows), as the appearance (captured by gray) is the most indicative cue for gait, while motion (captured by OF) can be subtle and not always discriminate. Nevertheless, in most cases our model results in more than 80% accuracy, highlighting that even with the absence of modalities, it successfully recognizes gait. Overall, using all modalities (last row) reaches the highest performance. For the multiview setup of CASIA-B, we report results at the video level in Tab. 5, after averaging on the 11 views. Gray achieves better accuracy than OF when the shape of the subjects is not altered drastically (i.e. NM and BG), but OF is more robust when clothes change (CL). Finally, the best results are obtained when using both modalities (last row).

4.5. Comparison to the state of the art

Tab. 1 reports the results of our method compared to the state of the art on TUM-GAID. Our methods outperforms all models on TUM-GAID. Specifically, on the elapsed-time scenarios (*TN, TB, TS*) the improvement is remarkable, reaching a mean accuracy of 98.8%. Fig. 2 depicts successful examples.

5. CONCLUSIONS

We introduced a novel framework that handles and combines various types of input modalities for gait recognition: gray, optical flow, and depth maps. Although it is trained with multiple modalities, at test time it is robust to missing ones. To the best of our knowledge, this is the first framework that enables gait recognition with missing modalities at test time. Our model sets the new state of the art on TUM-GAID.

6. REFERENCES

[1] W. Zeng, C. Wang, and F. Yang, "Silhouette-based gait recognition via deterministic learning," *PR*, 2014.

[2] Z. Zhang, L. Tran, X. Yin, Y. Atoum, X. Liu, J. Wan, and N. Wang, "Gait recognition via disentangled representation learning," in *CVPR*, 2019.

[3] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, "A comprehensive study on cross-view gait based human identification with deep CNNs," *IEEE PAMI*, 2017.

[4] H. Chao, Y. He, J. Zhang, and J. Feng, "Gaitset: Regarding gait as a set for cross-view gait recognition," in *AAAI*, 2019.

[5] F. M. Castro, M. J. Marín-Jiménez, N. Guil, and N. Pérez de la Blanca, "Automatic learning of gait signatures for people identification," in *IWANN*, 2017.

[6] C. Fan, Y. Peng, C. Cao, Xu Liu, S. Hou, J. Chi, Y. Huang, Q. Li, and Z. He, "Gaitpart: Temporal part-based model for gait recognition," in *CVPR*, 2020.

[7] C. Godard, O. Mac Aodha, M. Firman, and G. J Brostow, "Digging into self-supervised monocular depth estimation," in *ICCV*, 2019.

[8] A. Ranjan and M. J Black, "Optical flow estimation using a spatial pyramid network," in *CVPR*, 2017.

[9] M. Hofmann, J. Geiger, S. Bachmann, B. Schuller, and G. Rigoll, "The tum gait from audio, image and depth (gaid) database: Multimodal recognition of subjects and traits," *J. of Vis. Comm. and Image Repres.*, 2014.

[10] F. M Castro, M J Marín-Jiménez, and N Guil, "Multimodal features fusion for gait, gender and shoes recognition," *Machine Vision and Applications*, 2016.

[11] F. M. Castro, M. J. Marín-Jiménez, N. Guil, and N. Pérez de la Blanca, "Multimodal feature fusion for CNN-based gait recognition: an empirical comparison," *Neural Computing and Applications*, 2020.

[12] S. Choi, S. Lee, Y. Kim, T. Kim, and C. Kim, "Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared person re-identification," *CVPR*, 2020.

[13] C. Wan, L. Wang, and V. V Phoha, "A survey on gait recognition," *ACM Computing Surveys (CSUR)*, 2019.

[14] W. Sheng and X. Li, "Siamese denoising autoencoders for joints trajectories reconstruction and robust gait recognition," *Neurocomputing*, 2020.

[15] R. Liao, S. Yu, W. An, and Y. Huang, "A model-based gait recognition method with body pose and human prior knowledge," *PR*, 2020.

[16] M.M. Hasan and H.A. Mustafa, "Multi-level feature fusion for robust pose-based gait recognition using RNN," *IJCSIS*, 2020.

[17] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi, "Gait recognition using a view transformation model in the frequency domain," in *ECCV*, 2006.

[18] K. Bashir, T. Xiang, and S. Gong, "Gait recognition using gait entropy image," in *IET ICDP*, 2009.

[19] T. HW Lam, KH Cheung, and J NK Liu, "Gait flow image: A silhouette-based gait representation for human identification," *PR*, 2011.

[20] T. Randhavane, U. Bhattacharya, K. Kapsaskis, K. Gray, A. Bera, and D. Manocha, "The liar's walk: Detecting deception with gait and gesture," *arXiv*, 2020.

[21] L. Tran, X. Liu, J. Zhou, and R. Jin, "Missing modalities imputation via cascaded residual autoencoder," in *CVPR*, 2017.

[22] C. Wang, M. Niepert, and H. Li, "LRMM: Learning to recommend with missing modalities," in *EMNLP*, 2018.

[23] L. Wang, C. Gao, L. Yang, Y. Zhao, W. Zuo, and D. Meng, "PM-GANs: discriminative representation learning for action recognition using partial-modalities," in *ECCV*, 2018.

[24] A. Miech, I. Laptev, and J. Sivic, "Learning a text-video embedding from incomplete and heterogeneous data," *arXiv*, 2018.

[25] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NeurIPS*, 2014.

[26] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015.

[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NeurIPS*, 2015.

[28] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Proc. ICPR*, 2006.

[29] F. M. Castro, M.J. Marín-Jiménez, N. Guil Mata, and R. Muñoz Salinas, "Fisher motion descriptor for multiview gait recognition," *IJPRAI*, 2017.

[30] M. J. Marín-Jiménez, F. M. Castro, N. Guil, F. de la Torre, and R. Medina-Carnicer, "Deep multi-task learning for gait-based biometrics," in *ICIP*, 2017.