

Self-supervised Multi-Modal Video Forgery Attack Detection

Chenhui Zhao* and Xiang Li[†] and Rabih Younes[‡]

*Department of Electrical and Computer Engineering, University of Michigan, Ann Arbor, USA

[†]Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, USA

[‡]Department of Electrical and Computer Engineering, Duke University, Durham, USA

Abstract— Video forgery attacks threaten surveillance systems by replacing the video captures with unrealistic synthesis, which can be powered by the latest augmented reality and virtual reality technologies. From the machine perception aspect, visual objects often have RF signatures that are naturally synchronized with them during recording. In contrast to video captures, the RF signatures are more difficult to attack given their concealed and ubiquitous nature. In this work, we investigate multimodal video forgery attack detection methods using both visual and wireless modalities. Since wireless signal-based human perception is environmentally sensitive, we propose a self-supervised training strategy to enable the system to work without external annotation and thus adapt to different environments. Our method achieves a perfect human detection accuracy and a high forgery attack detection accuracy of 94.38% which is comparable with supervised methods. The code is publicly available at: <https://github.com/ChuiZhao/Secure-Mask.git>

Index Terms—Human Perception, Wireless Signal, Forgery Attack Detection

I. INTRODUCTION

In recent years, the unique properties e.g. concealment, penetration, and ubiquity have been extensively investigated in wireless-based perception methods. Person-in-WiFi [1] is a pioneering work attempting to address fine-grained human perception problems by using WiFi signals. The follow-up works Secure-Pose [2] and its improved version [3] propose learning-based methods to detect video forgery attacks using radio-frequency (RF) signals. Besides the video forgery attack task, RF-based methods can also achieve comparable performance against visual-based methods in other visual representation tasks [1], [4], [5]. However, most RF-based human perception methods are environmentally sensitive [1], [5]–[8], thus it is hard to adapt well-trained models to unseen environments, which severely prevents them from practical applications.

To date, several methods have been tried to tackle the adaptation problem. Person-in-WiFi [1] proposes a style-transfer method for CSI measurements utilizing the Cycle-GAN [9], but the performance gain of the proposed module is limited even after complicated synthetic data generation. WiPose [4] extracted the environmental weak-dependent Body-coordinate velocity profile(BVP) from CSI measurement combing with an antenna selection strategy to ease the influence of the background environment. Although many attempts have been made to mitigate the impact of environment changes [1], [4], [5], the cross-environment adaptation remain unfeasible since the RF-based human perception heavily depends on

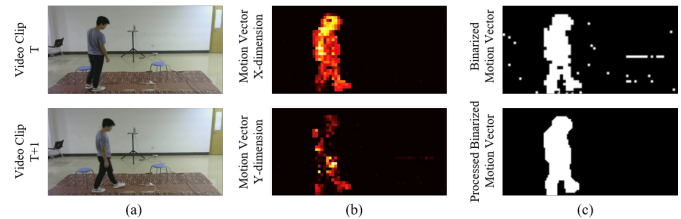


Fig. 1. Illustration of an original 2D motion vector and a binarized motion vector produced by our method from compressed video. The top and bottom figures of (a) are two adjacent frames in the video. The top and bottom figures of (b) are the X-dimension motion vector and the Y-dimension motion vector. The top and bottom figures of (c) are the original binarized motion vector and the binarized motion vector after processing.

the Doppler effect and electromagnetic property differences between human and background environment. Consequently, the gaps between RF-based human perception performance in the seen and unseen environment are hard to bridge from the perspective of data augmentation and denoising.

On the other hand, a surveillance camera is a periphery device for surveillance systems which captures visual information of target environments and transmits it to a central server. Due to the redundancy of the video modality, the video compression is always conducted before transmission. Motion vector is one of the essential components of the compressed video which reflects the block-wise spatial position changes across frames, and *de facto* fits the format of the label of the RF-based perception system.

In this work, to solve the adaptation problem of RF-based perception method, we introduce a self-supervised learning scheme to enable the model to learn from compressed video streams and further leverage it to conduct video forgery detection. We name the proposed system Secure-Mask. In particular, we adopt motion vectors from compressed video streams to create the supervision for RF-based model. Fig. 1 shows an example of motion vector and generated mask. Compared with other frame-based video forgery detection methods [10], [11], Secure-Mask can work in a real-time manner to generate fine-grained human segmentation and detect video replacement attacks at the object level. The concealed and ubiquitous properties of WiFi signals make it a good alternative to surveillance video and can be suitable for future secure systems to act on when cameras are offline, occluded or attacked. Moreover, the self-supervised training scheme enables Secure-Mask to adapt to the new environment which

eliminates the redundant data labeling after environmental changes (i.e. furniture movement). The contributions of this paper are as follows.

- We propose a self-supervised learning scheme for RF-based human perceptions leveraging the motion vector as a source of supervision and proving its ability to work without external annotations.
- We built up a self-supervised video forgery attack detection pipeline that can act in a real-time manner with high accuracy of 94.38%. The performance of Secure-Mask is comparable to its supervised counterpart Secure-Pose [2].

II. RELATED WORKS

Camera-based human perception. In the computer vision field, many works [12], [13] used well-developed feature extraction methods to accomplish challenging human perception tasks. In addition, there have been many works in object segmentation [14]–[18], pose estimation [19], [20], and activity recognition [21], [22]. More recently, depth cues obtained by the RGB-D camera have been introduced in human perception tasks and some other works [23], [24] have shown that depth cure can improve performance.

Sensor-based human perception. The Frequency Modulated Continuous Wave (FMCW) radar system was first introduced by Adib et al. [25] to capture coarse human bodies with a delicate radar device. Later, they extended this system to do pose estimation through the wall or other occlusions, with 2D [26] and 3D [27] included. Compared with the above methods, which rely heavily on expensive Radar equipment, WiFi signals provide a more ubiquitous and cheap option. However, WiFi-based works [6], [7] were not popularized before because they have not been producing fine-grained human masks or human skeletons until Wang et al. proposed Person-in-WiFi [1]. After that, more and more researchers paid attention to WiFi-based human perception works. For example, WiTA [8] recognized human activity in an attention-based way using commercial WiFi devices. Wi-Pose [5] reconstructed fine-grained human poses using WiFi signals. In addition, some previous works [28], [29] also investigate how to augment data to achieve better performance.

Video Forgery Detection. Forgery detection for surveillance systems has drawn researchers’ attention due to the advanced video forgery technologies. Many researchers solved this problem by analyzing the spatiotemporal features in surveillance video [10], [11]. These methods can determine the frame-based forgery, for example, frame delete and insert. For the object-based video forgery, Mohammed et al. proposed a sequential and patch analysis method [30], which can generate coarse forgery traces in each surveillance video frame. Relatedly, SurFi [31] compared timing information from WiFi signals and the corresponding live video to detect camera looping attacks. To generate fine-grained forgery traces while detection, Secure-Pose [2] first proposed a cross-modal system that can detect and localize forgery attacks in each video frame through a supervised way.

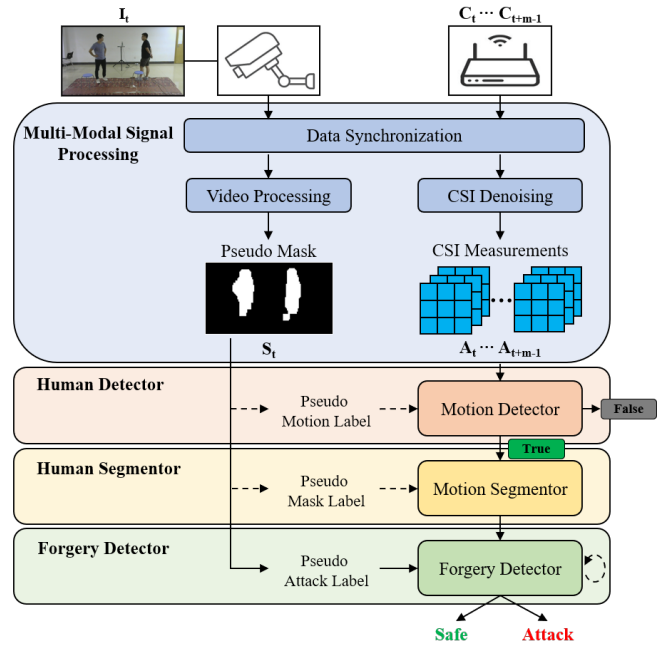


Fig. 2. Secure-Mask overview. Dash lines indicate model updating. Secure-Mask can be boiled down to four modules: multi-modal signal processing, human detector, human segmentor, and forgery detector. In the training phase, the motion vector M_t of video frame I_t is first synchronized to m CSI measurements $\{C_t, \dots, C_{t+m}\}$. Then we conduct the preprocessing separately for video and CSI data to get binarized masks S_t and CSI measurements in matrix form $\{A_t, \dots, A_{t+m}\}$. After that, we generate annotations for training the human detector and human segmentor networks. Finally, we use the predicted masks acquired from the human segmentor to generate annotations for training the forgery detector.

III. SECURE-MASK SYSTEM

In this section, we will elaborate on the detailed pipeline of our proposed Secure-Mask system.

A. System Overview

The Secure-Mask is composed of four parts: multi-modal signal processing, human detector, human segmentor, and forgery detector. As shown in Fig. 2, in the training phase, we leverage the motion vector in the live video stream to update the networks. In the multi-modal signal processing module, we first extract the motion vector $\{M_t\}_{t=0}^T$ from video frames $\{I_t\}_{t=0}^T$ and synchronize it with the CSI measurements $\{C_t\}_{t=0}^T$, then conduct signal processing separately. After that, the processed masks serve as labels to update networks. In the inference phase, only the CSI branch in the multi-modal signal processing module is activated. The processed CSI measurements are sent to the human detector to detect motions. To improve the efficiency of the whole system, the human segmentor and forgery detector only acts when the human detector confirms moving objects.

B. Multi-Modal Signal Processing

Secure-Mask is a cross-modal system that takes advantage of both visual and wireless modalities. To ensure both modalities contain homogeneous information, the data synchronization is essential. Given the nature of wireless communication,

wireless signals always have a much higher sample rate than video frames. Therefore, we assign multiple wireless frames to one video frame. Let us denote the captured video frames as $\{I_t\}_{t=0}^T$ and CSI measurements $\{C_t\}_{t=0}^T$ where $I_t \in \mathbb{R}^{3 \times H \times W}$ and $C_t \in \mathbb{C}^{K \times N_{tx} \times N_{rx}}$. H and W are the height and width of the video frames. N_{tx} , N_{rx} and K are the number of transmitters, receivers and subcarriers, respectively. We assign m CSI measurements $\{C_t, \dots, C_{t+m-1}\}$ to one frame I_t . We consider the amplitude of CSI measurements $\{A_t\}_{t=0}^T$ to make it a real matrix. After that we get the data pairs $\{I_t, \{A_t, \dots, A_{t+m}\}\}_{t=0}^T$.

CSI Processing. The environment noise can cause sudden changes in the CSI measurements, which will impact the efficiency of extracting the amplitude features from it. To filter out outliers in the CSI measurements, we utilize the Hampel identifier [32] to denoise the CSI data as Secure-Pose [2] did.

Video Processing. To improve video storage and transmission, it is common to perform video compression. Typically, the compression techniques such as MPEG-4 and H.264 leverage the temporal continuity of successive frames and retain only a few complete frames while reconstructing other frames using the motion vector and residual error. Our solution utilizes the 2D motion vector to create masks, which can be separated into binarization, denoising, and refinement. The motion vectors within a group of pictures (GOP) can be denoted as $\{M_t\}_{t=0}^G$ where $M \in \mathbb{R}^{H \times W \times 2}$. Since the velocity of human activity is slow compared to the video frame rate, by using a short GOP length, the human movement can be assumed as the same in each GOP. We determine the binary mask of the human movement from two dimensions of the motion vector, angle and amplitude. Let \hat{M} be the sum of all motion vectors in a GOP and \bar{M} be the $\phi_{3 \times 3}(\hat{M})$ where $\phi_{3 \times 3}$ is a gaussian smooth function with a kernel size of 3. For a single GOP, the binarized mask can be denoted as S and computed as:

$$S = \begin{cases} 1, & \text{if } \|\hat{M}_{i,j}\|_2 + \lambda \frac{\langle \bar{M}_{i,j}, \hat{M}_{i,j} \rangle}{|\bar{M}_{i,j}| \cdot |\hat{M}_{i,j}|} \geq \tau \\ 0, & \text{else} \end{cases} \quad (1)$$

where λ and τ are constants. The first term and second term filter motion vector based on amplitude and degree respectively. Here we set $\lambda = 1$ and $\tau = 0.5$. The mask S is further processed through a stack of soothing and morphological operations after binarization.

C. Human Detector

The human detector network is a lightweight network that aims to determine the existence of human motion. The Human detector takes CSI as input and outputs the binary result judging the human movement. We utilize both the convolution layer and the Long-Short Term Memory (LSTM) layer to process the CSI data, as it includes both spatial and temporal features. In particular, the CSI measurements $\{A_t, \dots, A_{t+m}\}$ are concatenated in the subcarriers dimension to form $A_t^c \in \mathbb{R}^{mK \times N_{tx} \times N_{rx}}$. The CSI data A_t^c is first processed by two convolution layers followed by one LSTM layer. The final

output is obtained by applying two linear layers to the LSTM output. With the human detector, we can save computational resources by only activating the following modules after detecting human motion.

We supervise the human detector by binary cross-entropy loss. Since the human detector is a binary classification network, the binarized ground-truth is obtained by the following criterion.

$$C(S) = \begin{cases} 1 & \text{if } \frac{\sum_{i=0}^W \sum_{j=0}^H S_{i,j}}{W \times H} \geq \eta \\ 0 & \text{else} \end{cases} \quad (2)$$

where W and H are the width and height of the pseudo mask S and η is a constant threshold, and we set $\eta = 0$.

D. Human Segmentor

Given the concatenated CSI data A_t^c , the human segmentor generates masks of moving humans in the perception field. To conduct this challenging task, a UNet-like structure is leveraged. The input CSI data A_t^c is first tiled to image size before feeding to the network. After that, the upsampled tensor is fed into an encoder to produce the encoded feature map. Then a transposed convolution-based decoder is utilized to transform wireless features to image space. Both encoder and decoder contain four downsample and upsample operations with a stride of 2 and, after each scaling operation, a 2D convolution is involved to refine the feature map before the next scaling. In particular, skip connections are used to retain swallow wireless features in later layers. Let us denote the human segmentor as \mathcal{S} , the human mask prediction $P_t \in \mathbb{R}^{H \times W}$ can be denoted as $P_t = \mathcal{S}(A_t^c)$.

We supervise the human segmentor by binary cross-entropy loss \mathcal{L}_{bce} and Dice loss \mathcal{L}_{Dice} . The overall loss for training is $\mathcal{L} = \mathcal{L}_{Dice} + \lambda_b \mathcal{L}_{bce}$.

E. Forgery Detector

The forgery detector aims to detect video forgery attacks using multimodal data. Since wireless perception heavily depends on the Doppler effect which is caused by the motion of objects, the motion vector and wireless data contain homogeneous representations of the moving humans in the perception area. We leverage the human mask as a proxy to conduct contrastive learning for video forgery detection. Given a clip of video $\{I_t, \dots, I_{t+g}\}$, we can obtain the motion vectors $\{M_t, \dots, M_{t+g}\}$ freely from the compressed video streams. We further processed it to obtain the pseudo mask $\{S_t, \dots, S_{t+g}\}$. The human segmentor predicts the human masks $\{P_t, \dots, P_{t+g}\}$ from the wireless modality. We tailor a network to compare the human masks from visual and wireless modalities to determine their homogeneity. The masks from the visual modality $\{S_t, \dots, S_{t+g}\}$ and masks from the wireless modality $\{P_t, \dots, P_{t+g}\}$ are concatenated as $\{Q_t, \dots, Q_{t+g}\}$ where $Q_t \in \mathbb{R}^{2 \times H \times W}$. We extract features for each time step separately with a ResNet-based network. After that, extracted features are fed into a one-layer LSTM

followed by two fully connected layers before the final output. We predict the homogeneity in a clip-wise manner.

Similar to the human detector, we utilize binary cross-entropy to supervise the training. We generate unmatched input sequence pairs to synthesis video forgery attacks by selecting unsynchronized masks pairs.

IV. EXPERIMENTS

To the best of our knowledge, there is no public multi-modal dataset for video forgery. Thus, we conduct experiments using the same dataset used in Secure-Pose [2]. In this section, we will show the qualitative result for the human segmentor and the quantitative result for both the human detector and forgery detector. In the quantitative result, we report the evaluation metrics including the accuracy (Acc), the false positive rate (FPR), and the true positive rate (TPR).

A. Dataset Description

The dataset (same as the dataset used in [2]) was collected in an $8m \times 16m$ office room with 5 volunteers. As shown in Table I, during the experiment phase, zero to three volunteers were asked to perform walking, sitting, waving hands, or random movements concurrently in the perception area.

TABLE I
STATISTIC OF THE DATASET. P: NUMBER OF CONCURRENT PERSON. F: NUMBER OF VIDEO FRAMES.

P	0	1	2	3	total
F	2242	4488	4498	911	12139

Human Detector. For the human detector model, we utilize all the video frames and their corresponding CSI measurements. Then we split them randomly, in which 9663 data pairs are used for training and 1024 data pairs are used for testing.

Human Segmentor. If the human in the video is not moving, we cannot leverage the motion vector to generate a reliable human mask. Therefore, we select those video frames that can generate valid motion vector and their corresponding CSI measurements. Then we split them randomly to train the human segmentor model first, in which 8574 data pairs are used for training and 870 data pairs are used for testing.

Forgery Detector. After the human segmentor model is well trained, we feed the CSI measurements from the human segmentor’s dataset into the human segmentor model to prepare the dataset for the forgery detector model. We get 8574 predicted masks and 870 predicted masks for the training and testing sets of the forgery detector, respectively. Then, we concatenate those predicted masks with the motion vector masks to get the labels: 0 if corresponding, else 1. Finally, for the forgery detector model, 8530 data pairs are used for training and 826 data pairs are used for testing.

B. Implementation Details

The human detector network, human segmentor network, forgery detector network are all implemented for 20 epochs on the Pytorch framework.

TABLE II
QUANTITATIVE RESULTS OF HUMAN DETECTOR AND FORGERY DETECTOR. WE SET $m = 5$ AND $g = 7$ FOR BOTH RESULTS.

Module	Acc	FPR	TPR
Human Detector	100%	0%	100%
Forgery Detector	94.38%	4.01%	92.27%

Human Detector. The learning rate starts from $1e - 6$ and is divided by 10 for each 5 epochs. The $batchsize = 16$ and a RMSprop optimizer [33] with $weight\ decay = 1e - 8$, $momentum = 0.9$ is leveraged.

Human Segmentor. The learning rate starts from $1e - 3$ and is divided by 10 for each 5 epochs. The $batchsize = 32$ and an adam [34] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $weight\ decay = 1e - 5$ is leveraged.

Forgery Detector. The learning rate starts from $1e - 3$ and is divided by 10 for each 5 epochs. The $batchsize = 32$ and an adam [34] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $weight\ decay = 2e - 5$ is leveraged.

C. Main Results

In this section, we report the performance of the human detector, human segmentor and the forgery detector. In addition, we also compare the result of the forgery detector with other recent methods.

Qualitative results. Since the dataset does not contain manually annotated segmentation labels, we only show the qualitative results of the predicted masks. As shown in Fig. 3, the predicted masks have the corrected spatial position, but the detailed shapes are not recovered. Two reasons can account for the shape difference. First, the spatial resolution of the commercial Wi-Fi signals is less than one decimeter, which makes it difficult for wireless data to capture detailed human boundary information. Second, the motion vector only provides coarse supervision compared to masks predicted by neural networks or manual annotation.

Quantitative results. We report the quantitative results of the human detector and the forgery detector. In the experiment, we set the number of CSI measurements per video frame $m = 5$ and the number of input video frames to the forgery detector $g = 7$. As shown in Table II, the human detector has a perfect performance. This is because the wireless signals are sensitive to moving objects. In the indoor scenario, the motion information can be effectively carried by the Wi-Fi signals: the CSI data contains the spatial information because of the Doppler effect, and as a kind of sampled signal, CSI data contains the temporal information naturally. Therefore, when considering the feature extraction strategy, combining the convolution layer with the LSTM layer is much better than the pure convolution operation as well.

The forgery detector also shows its promising performance, whose overall accuracy can reach 94.38%. We compare our method with other recent approaches [2], [11], [30], [31], which are supervised learning-based method. As shown in Table III, event- and frame-based forgery detection generally

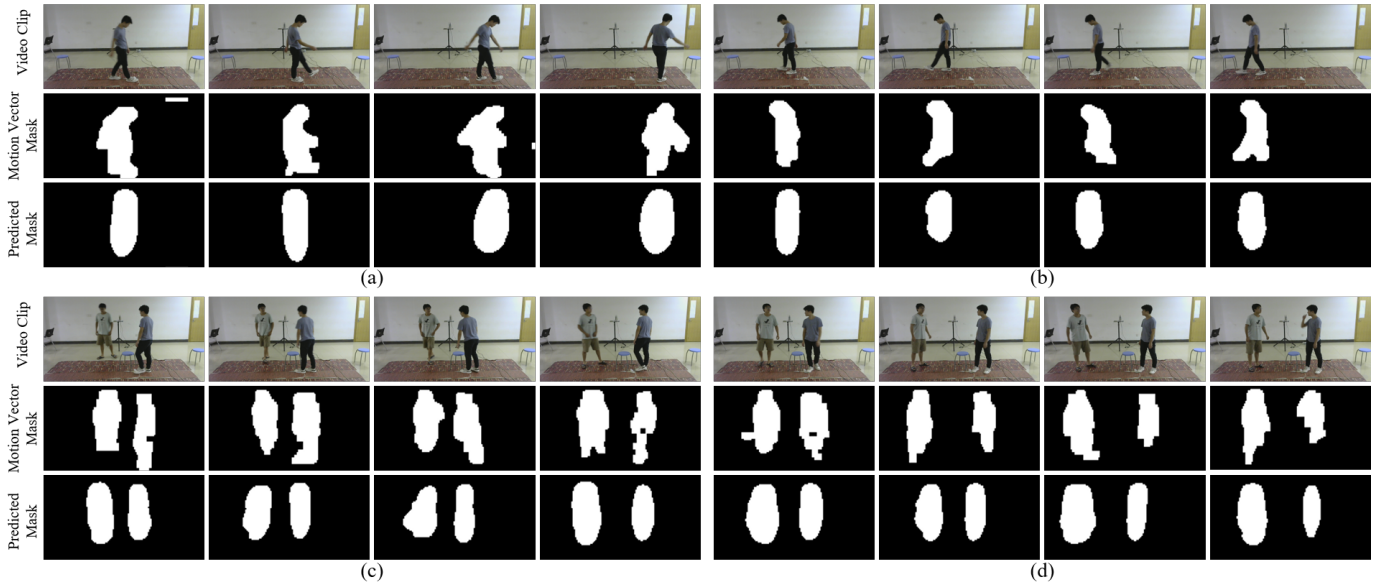


Fig. 3. Qualitative results of the human segmentor. We provide four video clips in total and each row contains two video clips. In clips (a) and (b), there is one person in the perception field and there are two people in the perception field for clips (c) and (d). For each clip, we show both the motion vector masks and the predicted masks, which reports the predicted masks having the corrected spatial position while the detailed shapes are lost.

TABLE III

COMPARISON RESULT OF FORGERY ATTACK DETECTION METHODS. THOSE METHODS USE DIFFERENT DATASETS TO INSTRUCT THE FORGERY ATTACK DETECTION AT DIFFERENT LEVELS. F: FRAME-BASED FORGERY DETECTION. E: EVENT-BASED FORGERY DETECTION. O: OBJECT-BASED FORGERY DETECTION

Method	Level	Multi-Modal	Acc
Fadl [11]	F		98.0%
Lakshmanan [31]	E	✓	98.9%
Aloraini [30]	O		93.18%
Huang [2] (Secure-Pose)	O	✓	94.9%
Secure-Mask (Ours)	O	✓	94.38%

have a higher accuracy than those at the object level. However, forgery detection only at the event and frame level will be limited in some situations (i.e., the forgery attack on the entire video). When considering the working pattern, the overall accuracy of Secure-Mask is only 0.52% lower compared to its counterpart, Secure-Pose, which needs work in a supervised way. However, our self-supervised approach enables the system to maintain its performance as the environment changes.

We also report the inference speed of the proposed system. As shown in Table IV, even using a normal GPU, our system can perform in a real-time manner.

D. Ablation Study

In this section, we conduct extensive ablation experiments to study the core factors of our method.

CSI measurements per video frame m . We first train the human segmentor with different amounts of the CSI measurements per video frame to evaluate the influence of the single predicted video frame on the forgery detector. As reported in Table V, with the number of the CSI measurements per video frame varying from 1 to 5, the Accuracy increases from 90.25% to 94.38%. This result suggests that an additional temporal information can make it easier for the human

TABLE IV

INFERENCE SPEED. ALL RESULTS ARE MEASURED ON SINGLE NVIDIA 2070 GPU. HD: HUMAN DETECTOR. HS: HUMAN SEGMENTOR. FD: FORGERY DETECTOR.

Module	HD	HS	FD
FPS	230	70	280

TABLE V

CSI MEASUREMENTS PER VIDEO FRAME. THE PERFORMANCE IMPROVE AS THE AMOUNT OF THE CSI MEASUREMENTS PER VIDEO FRAME INCREASES. THE RATIO HERE REFERS TO THE CSI MEASUREMENTS PER VIDEO FRAME.

Ratio (m)	Acc	FPR	TPR
1	90.25%	10.11%	90.57%
3	91.25%	3.47%	84.60%
5	94.38%	4.01%	92.27%

segmentor to learn how to decode human motion information from CSI measurements. Indeed, this kind of improvement makes the performance of the forgery detector better.

Number of the input video frames to forgery detector g . We then train the forgery detector with different numbers of the input video frames to evaluate the effect of the amount of the video frames on the model. As reported in Table VI, with the number varying from 3 to 7, the accuracy increases from 87.17% to 94.38%. This result shows that feeding more video frames into the model does improve its performance. The LSTM component in the forgery detector model can account for this since the LSTM component can learn the forgery information more efficiently when additional special information combined with temporal information is provided. As the video was recorded at 7.5 FPS from the camera, we set the largest amount of the input video frames as 7 to avoid fail detection with the short-time forgery. Moreover, we argue that if we recorded the video at a higher FPS, the forgery detector

TABLE VI
NUMBER OF THE INPUT VIDEO FRAMES. THE PERFORMANCE IMPROVE
AS THE NUMBER OF THE INPUT VIDEO FRAMES INCREASES

Frames (g)	Acc	FPR	TPR
3	87.17%	4.35%	79.08%
5	92.01%	4.39%	87.50%
7	94.38%	4.01%	92.27%

can achieve even better performance.

V. CONCLUSION

In this paper, we build a novel self-supervised system for video forgery detection eliminating the need for external annotations. Notably, our method achieves comparable performance against the previous supervised methods in forgery detection.

REFERENCES

- [1] Fei Wang, Sanping Zhou, Stanislav Panev, Jinsong Han, and Dong Huang, *Person-in-WiFi: Fine-Grained Person Perception Using WiFi*, pp. 1–14, Proceedings of the 26th Annual International Conference on Mobile Computing and Networking, 2019.
- [2] Yong Huang, Xiang Li, Wei Wang, Tao Jiang, and Qian Zhang, “Towards cross-modal forgery detection and localization on live surveillance videos,” in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 2021, pp. 1–10.
- [3] Yong Huang, Xiang Li, Wei Wang, Tao Jiang, and Qian Zhang, “Forgery attack detection in surveillance video streams using wi-fi channel state information,” *IEEE Transactions on Wireless Communications*, 2021.
- [4] Wenjun Jiang, Hongfei Xue, Chenglin Miao, Shiyang Wang, Sen Lin, Chong Tian, Srinivasan Murali, Haochen Hu, Zhi Sun, and Lu Su, “Towards 3d human pose construction using wifi,” in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, New York, NY, USA, 2020, pp. 1–14, Association for Computing Machinery.
- [5] Lingchao Guo, Zhaoming Lu, Xiangming Wen, Shuang Zhou, and Zijun Han, “From signal to image: Capturing fine-grained human poses with commodity wi-fi,” *IEEE Communications Letters*, vol. 24, no. 4, pp. 802–806, 2019.
- [6] Fadel Adib and Dina Katabi, “See through walls with wifi!,” in *Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM*, 2013, pp. 75–86.
- [7] Donny Huang, Rajalakshmi Nandakumar, and Shyamnath Gollakota, “Feasibility and limits of wi-fi imaging,” in *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems*, 2014, pp. 266–279.
- [8] Xiaolong Yang, Ruoyu Cao, Mu Zhou, and Liangbo Xie, “Temporal-frequency attention-based human activity recognition using commercial wifi devices,” *IEEE Access*, vol. 8, pp. 137758–137769, 2020.
- [9] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.
- [10] Dai-Kyung Hyun, Min-Jeong Lee, Seung-Jin Ryu, Hae-Yeoun Lee, and Heung-Kyu Lee, “Forgery detection for surveillance video,” in *The Era of Interactive Media*, pp. 25–36. Springer, 2013.
- [11] Sondos Fadd, Qi Han, and Qiong Li, “Cnn spatiotemporal features and fusion for surveillance video forgery detection,” *Signal Processing: Image Communication*, vol. 90, pp. 116066, 2021.
- [12] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li, “Solo: Segmenting objects by locations,” in *European Conference on Computer Vision*. Springer, 2020, pp. 649–665.
- [13] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick, *Mask R-CNN*, pp. 2961–2969, Proceedings of the IEEE international conference on computer vision, 2017.
- [14] Xiang Li, Jinglu Wang, Xiao Li, and Yan Lu, “Hybrid instance-aware temporal fusion for online video instance segmentation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, pp. 1429–1437.
- [15] Xiang Li, Jinglu Wang, Xiao Li, and Yan Lu, “Video instance segmentation by instance flow assembly,” *IEEE Transactions on Multimedia*, 2022.
- [16] Xiang Li, Jinglu Wang, Xiaohao Xu, Bhiksha Raj, and Yan Lu, “Online video instance segmentation via robust context fusion,” *arXiv preprint arXiv:2207.05580*, 2022.
- [17] Xiang Li, Jinglu Wang, Xiaohao Xu, Xiao Li, Yan Lu, and Bhiksha Raj, “R² v2os: Robust referring video object segmentation via relational multimodal cycle consistency,” *arXiv preprint arXiv:2207.01203*, 2022.
- [18] Xiang Li, Haoyuan Cao, Shijie Zhao, Junlin Li, Li Zhang, and Bhiksha Raj, “Panoramic video salient object detection with ambisonic audio guidance,” *arXiv preprint arXiv:2211.14419*, 2022.
- [19] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang, “Deep high-resolution representation learning for human pose estimation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5693–5703.
- [20] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little, “A simple yet effective baseline for 3d human pose estimation,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2640–2649.
- [21] Earnest Paul Ijjina and Krishna Mohan Chalavadi, “Human action recognition in rgb-d videos using motion sequence information and deep learning,” *Pattern Recognition*, vol. 72, pp. 504–516, 2017.
- [22] Nuno C Garcia, Pietro Morerio, and Vittorio Murino, “Modality distillation with multiple stream networks for action recognition,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 103–118.
- [23] Cristina Palmero, Albert Clapés, Chris Bahnsen, Andreas Møgelmoose, Thomas B Moeslund, and Sergio Escalera, “Multi-modal rgb–depth–thermal human body segmentation,” *International Journal of Computer Vision*, vol. 118, no. 2, pp. 217–239, 2016.
- [24] Xiaoqin Zhou, Xiaofeng Liu, Aimin Jiang, Bin Yan, and Chenguang Yang, “Improving video segmentation by fusing depth cues and the visual background extractor (vibe) algorithm,” *Sensors*, vol. 17, no. 5, pp. 1177, 2017.
- [25] Fadel Adib, Chen-Yu Hsu, Hongzi Mao, Dina Katabi, and Frédo Durand, “Capturing the human figure through a wall,” *ACM Transactions on Graphics (TOG)*, vol. 34, no. 6, pp. 1–13, 2015.
- [26] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi, “Through-wall human pose estimation using radio signals,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7356–7365.
- [27] Mingmin Zhao, Yonglong Tian, Hang Zhao, Mohammad Abu Alsheikh, Tianhong Li, Rumen Hristov, Zachary Kabelac, Dina Katabi, and Antonio Torralba, “Rf-based 3d skeletons,” in *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, 2018, pp. 267–281.
- [28] Xi’ang Li, Jinqi Luo, and Rabih Younes, “Activitygan: Generative adversarial networks for data augmentation in sensor-based human activity recognition,” 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*, 2020, pp. 249–254.
- [29] Jinqi Luo, Xiang Li, and Rabih Younes, “Toward data augmentation and interpretation in sensor-based fine-grained hand activity recognition,” in *International Workshop on Deep Learning for Human Activity Recognition*. Springer, 2021, pp. 30–42.
- [30] Mohammed Aloraini, Mehdi Sharifzadeh, and Dan Schonfeld, “Sequential and patch analyses for object removal video forgery detection and localization,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 3, pp. 917–930, 2021.
- [31] Nitya Lakshmanan, Inkyu Bang, Min Suk Kang, Jun Han, and Jong Taek Lee, “Surfi: detecting surveillance camera looping attacks with wi-fi channel state information,” in *Proceedings of the 12th Conference on Security and Privacy in Wireless and Mobile Networks*, 2019, pp. 239–244.
- [32] Ronald K Pearson, “Outliers in process modeling and identification,” *IEEE Transactions on control systems technology*, vol. 10, no. 1, pp. 55–63, 2002.
- [33] Tijmen Tieleman, Geoffrey Hinton, et al., “Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude,” *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.
- [34] Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.