CRONOS: Colorization and Contrastive Learning Enhanced NLoS Human Presence Detection using Wi-Fi CSI Signals

Li-Hsiang Shen, Chia-Che Hsieh, An-Hung Hsiao, and Kai-Ten Feng

Department of Electronics and Electrical Engineering
National Yang Ming Chiao Tung University, Hsinchu, Taiwan

gp3xu4vu6.cm04g@nctu.edu.tw, chiachehsieh@gmail.com,
eric28200732.cm07g@nctu.edu.tw, and ktfeng@nycu.edu.tw

Abstract

In recent years, demands of pervasive smart services and applications increase explosively. Device-free human detection through sensors or cameras has been widely adopted but with privacy issues as well as misdetection for motionless people. To resolve these defects, channel state information (CSI) captured from commercialized Wi-Fi devices is capable of providing plentiful signal features for accurate detection. The existing systems has inaccurate classification under a non-line-of-sight (NLoS) and stationery scenario of a person standing still at corner in a room. In this work, we have proposed a colorization and contrastive learning enhanced NLoS human presence detection (CRONOS) system. CRONOS is capable of generating dynamic recurrence plots (RPs) and coloring CSI ratios to distinguish mobile people and vacancy of a room, respectively. Furthermore, supervised contrastive learning is conceived to retrieve substantial representations, where consultation loss is formulated to differentiate the representative distances between dynamic and stationery cases. Furthermore, a self-switched static feature enhanced classifier (S3FEC) is proposed to determine the utilization of either RPs or coloring CSI ratio. Finally, comprehensive experimental results have revealed that our proposed CRONOS outperforms the existing systems applying machine learning, non-learning based methods as well as non-CSI based features in open literature, which achieves the highest presence detection accuracy and moderate computational complexity in vacancy, mobility, LoS and NLoS scenarios.

Index Terms

Machine learning, deep learning, contrastive learning, non-line-of-sight (NLoS), human presence detection, recurrence plot, channel state information (CSI), colorization, consultation loss.
I. INTRODUCTION

With the popularity of smart homes, many intelligent applications have sprung up, such as smart lighting, media entertainment, and home energy management system. In addition to the pursuit of convenience and amusement, home security applications have also received considerable attention, involving elderly care [1], intrusion detection [2], and disaster prevention. For these applications, indoor human presence detection is a necessary and core underlying technology, especially the device-free scenarios where there is no need for targeting devices to be equipped with electronic devices. Nowadays, the most widespread device-free human presence detection technologies are adopting infrared sensors [3] and camera-based methods [4]. The advantage of the infrared sensor is that it is inexpensive and energy-saving. However, infrared sensor-based techniques only work when a person is moving in line-of-sight (LoS) and become futile for the non-line-of-sight (NLoS) or motionless scenarios. In addition, although camera-based methods can provide more comprehensive image information, they are impeded by privacy apprehensions and blindspot troubles. In contrast, Wi-Fi signals can not only pass through walls to avoid blindspot problems but also do not involve privacy issues. Therefore, Wi-Fi signal-based methods have become a prevalent research topic in human presence detection.

The channel state information (CSI) can be estimated from the Wi-Fi signal through channel sounding. The original functions of CSI are to represent channel characteristics and generate an appropriate equalizer from it to compensate for wireless channel effects. Nevertheless, since human activities also affect the wireless channel, CSI is widely utilized in wireless sensing tasks. Compared with received signal strength (RSS) [5], CSI is fine-grained information and can be decomposed into amplitude and phase information. Therefore, several wireless sensing tasks using CSI amplitude or phase information are proposed in recent years [6–13]. In [6] and [7], they employ deep learning methods to conduct fingerprint-based indoor localization after preprocessing amplitude and phase, respectively. The papers of [8] and [9] implement human counting based on the calibrated CSI phase. In the study of vital sign monitoring, the authors of [10] detect the vital sign and postures during sleep with CSI amplitude, and [11] uses the phase difference between receiving antennas to monitor respiration and heartbeat. However, the aforementioned papers only adopt the amplitude or phase information of CSI to carry out their tasks, which makes the detection ability of CSI weaker because of using incomplete CSI information. By contrast, the authors in [12] convert the amplitude and phase difference into images for indoor positioning. [13] employs the CSI ratio between the antennas to achieve respiration sensing with a wide detection range. They utilize the amplitude and phase information to improve the accuracy or expand the detection area. For indoor human presence detection, the detectable range must cover the entire indoor space. Otherwise, the person in the detection blindspot will lead to the incorrect judgment. Therefore, the CSI ratio is considered suitable for human presence detection since combining amplitude and phase information in the form of complex number can provide stronger detection capabilities. Converting the CSI ratio into the complex plane image can be treated as a novel design to represent the complete features of CSI ratio.
Human activities are related to time, which leads to CSI data possessing corresponding characteristics in time. Thus, in many wireless sensing studies, the time series data of CSI is used for classification [14], [15]. Furthermore, we can convert time series of CSI into images with temporal features, and recurrence plots (RPs) [16] are an example. RPs are diagrams that can analyze the recurrence in a time series, and we can identify from RP whether the signal is periodic or chaotic during a specific time period. RPs have been widely utilized in qualitative analysis by calculating some parameters in many fields of study. The paper [17] constructs RP from the radar sea clutter data to classify sea surface floating and sea clutter. In addition to the form of images, the recurrence rate is computed from RP to help their classification. In [18], after adopting gastric slow wave to create RP, they calculate the recurrence rate and diagonal line entropy from RP for gastrointestinal field analysis. In CSI-based wireless sensing tasks, the authors of [19] and [20] intend to implement a respiration detection system. In their designs, they generate RPs from all subcarrier values of CSI data and use the periodicity of RPs to perform subcarrier selection. With the prosperous research adopting RPs to represent temporal features in different fields, we intend to use RPs in presence detection to distinguish the cases between an empty room and the existence of a dynamic person under the environments with time-changing human behaviors. Moreover, with the assistance of RPs, we are able to develop powerful image classification algorithms to enhance accuracy for human presence detection.

After converting CSI into images, we require a learning method to classify these images, and contrastive learning has become very popular recently. In the beginning, SimCLR [21] proposed a simple framework for contrastive learning on visual representation. They employed a self-supervised learning algorithm for image recognition and achieved top-1 accuracy on ImageNet. After that, many papers adopt the architecture of self-supervised contrastive learning [22–24]. However, the self-supervised strategy implies that they do not utilize the complete label information. Therefore, SupCon [25] proposes supervised contrastive learning that can allow us to employ label information. Currently, although the cross-entropy loss is the most widely used for supervised classification problems, several papers have revealed that this loss has the shortcomings of lack of robustness to noisy labels [26, 27] and the possibility of poor margins [28, 29]. The results of SupCon also confirm that their supervised contrastive loss is better than cross-entropy loss on supervised learning. Some wireless sensing applications employ self-supervised or supervised contrastive learning architectures to implement their work. STF-CSL [30] adopts a self-supervised contrastive learning framework combined with time-domain and frequency-domain data augmentation to perform human activity recognition tasks. The paper [31] employs the architecture of supervised contrastive learning to complete fingerprint-based positioning on real-world outdoor CSI data. We can know that contrastive learning has been very successful in many fields, especially computer vision. For human presence detection, contrastive learning is a mighty classification algorithm when converting CSI into images to classify. Moreover, the presence detection adopting the learning method still needs to collect labeled data. Supervised contrastive learning can also be more capable of solving the problem of noisy labels.
than cross-entropy loss.

Recently, many studies using CSI signals for human presence detection have also been proposed. [32] adopts density-based spatial clustering of applications with noise to reduce the noise influence on CSI signals. Then they use support vector machines (SVM) to determine whether there is a walking or standing person in the room. In [33], they filter CSI data through Butterworth and moving average filter and use long short-term memory (LSTM) to classify whether there is someone in the indoor space. The authors of [34] transform both CSI amplitude and phase information into images. Afterward, two parallel convolutional neural networks (CNNs) are applied to extract the amplitude and phase features to classify the empty and walking human cases. The paper [35] utilizes their proposed convolutional denoising autoencoder to reduce the dimensionality of CSI data. After that, the processed CSI data will be inputted into the neural network to achieve multiple spot presence detection. In [36], they extract dynamic and spatial domain features in data preprocessing and design a conditional recurrent neural network for multi-room human presence detection. The above papers implemented human presence detection in many different situations, including a moving person, a stationary person, and a multi-room scenario. However, none of them mentioned the issue of NLoS standing person. Since the still person in the corner cannot cause a noticeable disturbance to the wireless channel, the signal characteristics of the CSI data are very similar to those of the empty room case. Therefore, some designs are essential for the NLoS static problem such that a human presence detection system can be realized in daily life.

As mentioned in the previous paragraphs, many effective designs can be applied to CSI-based human presence detection, such as CSI ratio, RPs, and contrastive learning. However, we are the first work to incorporate these techniques and actually apply them to human presence detection. In addition, those studies that use CSI signals for indoor human presence detection can not deal with the confusion problem between an empty room and a standing still person in corners. Therefore, we propose the colorization and contrastive learning enhanced NLoS human presence detection (CRONOS) system. The colorization algorithm for the CSI ratio image is designed to solve the NLoS static problem. Moreover, we also adopted RPs to help us judge whether the person is walking or not. Finally, supervised contrastive learning is also employed to learn representations to achieve a human presence detection system that can detect the stationary person in corners. The main contributions of this work are stated in the followings.

• Our proposed CRONOS system can implement human presence detection in a single-room scenario, which can classify an empty room, an NLoS standing person, an LoS standing person, and a walking person. To the best of our knowledge, we are the first to design a system capable of detecting the stationary person in NLoS corners. Other existing methods only deal with the person moving or static in places that can significantly affect the CSI signal, such as LoS. They all deliberately avoid the problem of the undetectable NLoS human.
• The CRONOS system consisted of feature-encompassed image generation (FEIG) and three-stage su-
supervised contrastive learning. FEIG can generate dynamic and static feature images. Dynamic feature (DF) images are RPs generated from the CSI amplitude difference between antennas and can be used to distinguish dynamic and static situations. The coloring CSI ratio images generated by the colorization algorithm are static feature (SF) images to separate static cases. Colorization can enlarge the visual difference between empty and NLoS cases to help resolve the NLoS static problem.

- After transforming the CSI data into images in FEIG, we propose three-stage supervised contrastive learning to classify these images into four cases of human presence detection. In stage 1, we employ supervised contrastive loss to train the representations of RPs. Stage 2 further adopts our designed consultation loss to learn the representations of coloring CSI ratio images. They are more capable of solving the NLoS static problem. In the last stage, we design the self-switched static feature enhanced classifier (S3FEC) to let the classifier automatically decide which representation to use for classification, which can solve the restriction problem between these two representations.

- We evaluate the performance of our CRONOS system in two real-world scenarios. It can be observed that CRONOS can achieve the highest F1-score in the NLoS standing case and the overall performance compared to other existing methods that did not have a specific design for an NLoS stationary person, demonstrating that our system can solve the NLoS static problem. Moreover, the ablation study confirmed that any design in our system has significant help to our system.

The rest of this thesis is organized as follows. Chapter II describes the system architecture of considered problem and preliminary of CSI, and the proposed CRONOS is explained in Chapter III. Chapter IV provides performance evaluation; while the conclusions are drawn in Chapter V.

II. System Architecture and Preliminary Observations

A. System Architecture

Our work aims to design a device-free human presence detection system that focuses on detecting whether there is a standing person in NLoS scenarios in an indoor environment. In addition to basically distinguishing the empty room from human presence cases, our system also categorizes three behaviors of the person in the room. Fig. I shows the system architecture of our CRONOS scheme. Two commercial off-the-shelf (COTS) Wi-Fi access points (APs) are deployed as the transmitter and receiver. Both APs operate in the multi-input multi-output (MIMO) and orthogonal frequency division multiplexing (OFDM) mechanisms in IEEE 802.11n. After collecting customized packets from the transmitter, receiver can estimate the CSI by some channel sounding methods. Note that human absence or presence in various scenarios leads to different multipath behaviors, which triggers the algorithm design for presence detection based on CSI signals. In our system, we will collect four cases of CSI data which are described as follows:

- case 1: empty room (static case),
Fig. 1: System architecture for indoor human presence detection.

- case 2: a standing still person in the NLoS scenarios (static case),
- case 3: a standing person in the LoS scenarios (static case),
- case 4: a walking person around the room (dynamic case).

Notice that we only consider these three cases of human presence because they take up almost situation in our daily life. At the same time, the receiver sends CSI data to edge nodes for data saving and processing. Finally, our CRONOS system can predict the detection results, and the results will be displayed on the monitor to make it easy to examine.

B. Channel State Information

In recent years, OFDM has become the dominant technique in wireless communication systems because of some advantages. For instance, OFDM is the approach that can tolerate multipath delay due to long symbol time and guard period. Besides, the MIMO system equipped with multiple antennas on the transmitter and receiver can provide a higher data rate and improve link quality. Hence, IEEE 802.11n Wi-Fi standard adopts the MIMO-OFDM system. Based on MIMO-OFDM system, the CSI data in one packet consist of multiple complex values that offer fine-grained information, and the quantity depends on the number of transmission pairs and OFDM subcarriers. These CSI signals represent the channel property influenced by some phenomena such as reflection, refraction, diffraction, and scattering. Thus, the furniture placement or human behavior makes
the CSI signal different in the indoor environment. As mentioned above, we can express the received CSI matrix at time \( t \) as

\[
H^t = \begin{bmatrix}
    h^t_{1,1} & \cdots & h^t_{1,n} & \cdots & h^t_{1,N} \\
    \vdots & \ddots & \vdots & \ddots & \vdots \\
    \vdots & \ddots & \vdots & \ddots & \vdots \\
    \vdots & \ddots & \vdots & \ddots & \vdots \\
    h^t_{m,1} & \cdots & h^t_{m,n} & \cdots & h^t_{m,N} \\
    \vdots & \ddots & \vdots & \ddots & \vdots \\
    \vdots & \ddots & \vdots & \ddots & \vdots \\
    \vdots & \ddots & \vdots & \ddots & \vdots \\
    h^t_{M,1} & \cdots & h^t_{M,n} & \cdots & h^t_{M,N}
\end{bmatrix},
\]

where \( m \) and \( M \) represent the index and total number of transmitter antennas, respectively, and \( n \) and \( N \) are the same relationships in the receiver. And the vector of the \( m \)-th transmitting antenna and the \( n \)-th receiving antenna means the CSI data of one transmission pair is defined as

\[
h^t_{m,n} = [h^t_{m,n,1}, h^t_{m,n,2}, \ldots, h^t_{m,n,k}, \ldots, h^t_{m,n,K}],
\]

where \( k \in [1, \ldots, K] \) is the subcarrier index with \( K \) as the maximum number of subcarriers. Each element of \( h^t_{m,n} \) is a superposition of all signal paths, which characterizes the multipath propagation in respective subcarriers. As a result, the CSI value can be modeled as

\[
h^t_{m,n,k} = e^{-j\phi^t_{m,n,k}} \sum_{l=1}^{L_{m,n}} A^t_{m,n,l} e^{-j2\pi \frac{d^t_{m,n,l}}{\lambda_k}} = |h^t_{m,n,k}| e^{j\angle h^t_{m,n,k}},
\]

where \( \phi^t_{m,n,k} \) is a random phase offset, and \( L_{m,n} \) is the total number of propagation paths that arrive at the \( n \)-th receiver antenna from the \( m \)-th transmitter antenna. \( A^t_{m,n,l} \) and \( d^t_{m,n,l} \) are respectively the complex attenuation and propagation length of the \( l \)-th path, and \( \lambda_k \) is the wavelength of the \( k \)-th subcarrier. Besides, \( |h^t_{m,n,k}| \) and \( \angle h^t_{m,n,k} \) denote the amplitude information and phase information of the transmission pair of the \( m \)-th transmitter and \( n \)-th receiver antenna at the \( k \)-th subcarrier, respectively. In most CSI sensing research, these two kinds of information are crucial components because they will change along with human activity, including amplitude attenuation and phase shift in the form of the real number. However, there is the random phase offset \( \phi^t_{m,n,k} \) due to the lack of perfect time or frequency synchronization in COTS Wi-Fi APs. This phase offset error makes it arduous to utilize the raw CSI phase information for sensing tasks. Nonetheless, we still intend to incorporate the merits of CSI phase data in our design for presence detection. We alternatively use the CSI ratio between antennas to avoid the useless raw phase problem. The explanation and derivation will be presented in the next chapter.

C. Observations of CSI Signals in Human Presence Detection

Since rich multipath signals reach the receiver in the indoor scenario, the CSI data can effectively reflect human activity and environmental change. Hence, the primary goal of our system is to differentiate the four
Fig. 2: Two CSI information of four cases, including (a) amplitude information and (b) unwrapped phase information.

cases by CSI signals. However, using CSI for classifying these four cases is not simple, especially in the NLoS standing case. For this reason, we conduct preliminary experiments to observe CSI amplitude and phase information within different scenarios, thereby realizing why the NLoS case is ambiguous and acquiring some concepts for designing proper approaches. The following results are exemplified observations from a transmission pair.

1) NLoS Static Problem: First, as shown in Fig. 2(a), we compared CSI amplitude information of four cases, where different lines in each case mean the 100 samples in continuous packets. The amplitude patterns in the empty room and NLoS standing case are very similar. The reason is that the different paths in these two situations can not impact the overall CSI amplitude value because these paths have severe attenuation. Furthermore, as cases 1 and 2 belong to static conditions, their CSI amplitudes do not enormously fluctuate on the time axis. As a result, it is challenging to distinguish these two cases by only using CSI amplitude information. In contrast, the shapes of CSI amplitude in the LoS standing case are vastly different from the empty room and NLoS case due to the blocked LoS path. And the CSI amplitude of human movement will undulate over time because the walking person causes the changes in the wireless channel. These influences let us can easily separate these two scenarios from other cases.

Next, we pay attention to the CSI phase information of four cases which are the same samples as the amplitude data. Note that we unwrapped the raw CSI phase along the time axis. As illustrated in Fig. 2(b), it is impractical to discriminate between them by only phase information, except that the LoS case has a visible difference. The main reason is that the CSI phase is contaminated with the random phase offset mentioned in equation (3). This offsets will make the raw CSI phase useless for sensing tasks.

In summary, if we only adopt the CSI amplitude or phase information for our work, it is laborious to classify the situations of human absence and person standing in the NLoS corner. That is why many researchers
deliberately avert the NLoS static problem in human presence detection work.

2) Dynamic Feature for The Walking Person: Other experimental results that we present are the CSI amplitude difference between two receiver antennas in four cases. As shown in Fig. 3, we can observe that the moving case has a huge fluctuation, but the values in the others are relatively smooth. The reason is that the empty room, NLoS standing person, and LoS standing person belong to static cases. The paths and attenuation of signals sent from the transmitter to the two receiver antennas do not vary with time. Hence, the amplitude difference in these three cases should be steady. On the other hand, the walking person results in different influences on signals in different timestamps. For instance, if the moving human blocks the LoS path of one antenna at the time one but makes it nonblocking at the time two, the CSI values will change a lot on two antennas between these two timestamps. Therefore, it will contribute to a noticeable oscillation in amplitude difference values. By this characteristic, we can design an approach to detect whether the person is dynamic or not in the room.

III. PROPOSED CRONOS SYSTEM MODEL

With the observation results of CSI data in the previous chapter, we comprehend that we can not merely employ the CSI amplitude or phase information to resolve the NLoS static problem. Additionally, the CSI amplitude difference between antennas can be processed to identify whether the indoor person is moving. Therefore, in this chapter, we will exhaustively describe how to deal with the NLoS static problem and discriminate the cases of human absence and three behaviors of human presence.

Fig. 4 depicts the block diagram of our CRONOS system for presence detection, and it can be divided into offline and online phases. The offline phase consists of two parts, inclusive of FEIG and three-stage supervised contrastive learning. In FEIG, our design is to transform the CSI data into images to turn our presence detection problem into an image recognition problem. Firstly, the DF extraction function extracts DF from a CSI time series of two transmission pairs. Then the DF image generation converts the DF into an image as the input to separate the dynamic case from static cases. Secondly, the SF extraction is conducted at the last time slot of the
Fig. 4: Block diagram for proposed CRONOS system.
CSI sequence. The SF contains multiple couples, and each one will generate a coloring image by colorization algorithm. After that, these images of couples will be merged as the input for distinguishing three static cases. Next, let us focus on the three-stage supervised contrastive learning. In stage 1 shown in the top-right of Fig. 4, supervised contrastive learning trains the contrastive encoder for the DF. The goal is to acquire the dynamic representations for four cases. In the middle-left part of Fig. 4 stage 2 concentrates on the contrastive encoder for the SF. It combines the consultation loss that can assist the model in getting better static representations. In stage 3 displayed in the middle-right of Fig. 4 the DF and SF encoders obtained from stages 1 and 2 are frozen and associated with S3FEC to do the feature map. After that, the weights of online model can be learned by computing the loss and implementing the backpropagation mechanism. In the online phase, we also transform the real-time CSI data into images by FEIG. Eventually, our system can predict the real-time results of presence detection by inputting the image data into the well-trained model. Based on the process, we can achieve high accuracy of presence detection even when a standstill person is in the NLoS corner.

The following subsections interpret how each block works and explain why they can help us accomplish the human presence detection system that focuses on the NLoS static scenario.

A. Feature-Encompassed Image Generation (FEIG)

In this section, we spell out each block of the FEIG. As mentioned above, our system converts the CSI data into images, e.g., DF image generation and SF image generation. Moreover, colorization and channel merging can reinforce the SF. Therefore, the FEIG is designed to generate images that encompass dynamic and static features of the CSI for human presence detection.

1) Dynamic Feature Extraction and Recurrence Plot: As shown in section II-C2 the amplitude difference in the human walking case has acute variations along the time axis; while the other three are relatively steady. We apply the amplitude difference as the DF in our work to differentiate between stationary and human motion situations. The amplitude difference of two transmission pairs from the \( m \)-th transmitter antenna at time \( t \) and \( k \)-th subcarrier can be expressed as

\[
d_{m,n_1,k}^t = |h_{m,n_1,k}^t| - |h_{m,n_2,k}^t|,
\]

where \( n_1, n_2 \in [1, N] \), \( n_1 \neq n_2 \), \( n_d = (n_1, n_2) \) represents the pair of two receiving antennas, and the subscript \( d \) means the difference. In addition, to make the characteristics for each case more conspicuous and neater, we average all subcarrier values, which can be calculated as

\[
\bar{d}_{m,n_d}^t = \frac{1}{K} \sum_{k=1}^{K} d_{m,n_d,k}^t.
\]
Next, we need to produce a time window as our DF since these features emerge only if we observe amplitude difference values within a time period, which can be represented as

$$f_d = [\bar{d}_{m,n}^{t-\tau+1}, \bar{d}_{m,n}^{t-\tau+2}, \ldots, \bar{d}_{m,n}^{t}]$$

(6)

where the $\tau$ is the window size. Combining with the results of section II-C2, the elements of $f_d$ will be capricious in the person moving case, and that in the other three static cases will be constant.

The next step is to turn the DF from numerical values into DF images that can enlarge the discrepancy between static and dynamic cases. For this purpose, RPs [16] are employed in our system as the DF image. RPs are the graphs that can examine whether a statistical value or system backs to the original state at some time. If the values are the stochastic signal, the points in RPs are more uniform. And if the values are the periodic signal, there are some regular structures in RPs. Our system directly adopts the RPs in image form as our classification input, and we can get RPs from DF computed in the equation (6). Notice that RP can be expressed as a matrix, and each element can be formulated as

$$x_{rp}^{t_1,t_2} = \begin{cases} 1 & |\bar{d}_{m,n}^{t_1} - \bar{d}_{m,n}^{t_2}| \leq \gamma \\ 0 & |\bar{d}_{m,n}^{t_1} - \bar{d}_{m,n}^{t_2}| > \gamma \\ \end{cases}$$

(7)

where $t_1, t_2 \in [t - \tau + 1, t]$, and the $\gamma$ is the threshold of RPs. If the absolute value of the difference between $\bar{d}_{m,n}^{t_1}$ and $\bar{d}_{m,n}^{t_2}$ is lower than $\gamma$, the element value is set as 1, denoting that the point in the image is black. Besides, if the absolute value is greater than $\gamma$, the pixel of the graph is white. In this work, the threshold $\gamma$ is based on the $\bar{d}_{m,n}$ of the empty case in a specific time interval. We define the candidate set of $\gamma$ as

$$D_e = \{|\bar{d}_{m,n}^{t_1}(e) - \bar{d}_{m,n}^{t_2}(e)| \leq \gamma | t_1, t_2 \in [t - \tau, t] \},$$

(8)

where $e$ means case 1 (empty room), and $\tau$ is the window size for the time interval. We choose the $\gamma = |\bar{d}_{m,n}^{t_1}(e) - \bar{d}_{m,n}^{t_2}(e)|$ in $D_e$ that makes cumulative distribution function (CDF) $F_{D_e}(\gamma) = 0.9$ so that the average value of amplitude difference in empty case is hardly more than $\gamma$. Hence, most of the elements in RPs in the human absence case will be black. And the NLoS and LoS cases have an equivalent result due to their similar average value of amplitude difference. Because the dynamic case has a perturbation in amplitude difference, it is reasonable to exceed the value of $\gamma$, which lets RP have white blocks. Thus, RPs can meet the objective of broadening the distinction between static and dynamic cases. Eventually, RPs can be represented as

$$X_{rp} = \begin{bmatrix} x_{rp}^{t-\tau+1,t-\tau+1} & x_{rp}^{t-\tau+1,t-\tau+2} & \ldots & x_{rp}^{t-\tau+1,t} \\ x_{rp}^{t-\tau+2,t-\tau+1} & x_{rp}^{t-\tau+2,t-\tau+2} & \ldots & x_{rp}^{t-\tau+2,t} \\ \vdots & \vdots & \ddots & \vdots \\ x_{rp}^{t,t-\tau+1} & x_{rp}^{t,t-\tau+2} & \ldots & x_{rp}^{t,t} \end{bmatrix}$$

(9)
Fig. 5 is a schematic diagram for realizing how RPs work and their properties. There are two time windows, where the window size $\tau = 5$. One is more fluctuating, and the other is relatively steady. Notice that the data below and left represent the same time window, and we set the threshold $\gamma = 0.3$. In Fig. 5(a) if we compute the absolute value of difference between time 1 and 3, the result equals 0.6 is greater than 0.3, so the block is white. In Fig. 5(b) we also calculate the absolute value of difference between time 1 and 3, but 0.2 is less than 0.3, which lets the block be black. After acquiring the results of all blocks, we can discover that RP is a persymmetric matrix, and the diagonal must be black because their differences are all equal to 0. Furthermore, the more fluctuating one has more white blocks, and the less one is darker. With this property, we can efficiently distinguish between dynamic and static cases by RPs.

2) Static Feature Extraction and Coloring CSI Ratio Image: Besides the DF, we will explain how to extract the SF and why our SF is better than only CSI amplitude or phase on NLoS static problem in this subsection. In addition, we design a colorization method to enhance the effects of our SF, and the channel merging can increase the data abundance but does not boost the complexity. The following three parts will provide detailed descriptions of each strategy.

a) CSI Ratio and Complex Plane Image: Many studies [35], [36] on human presence detection only adopt raw and processed CSI amplitude as the classification basis. However, as the observation results revealed in section II-C1, only applying CSI amplitude is difficult to discern the signal characteristics between empty room and NLoS static case. The insignificant changes in amplitude information results in the false classification of these two cases. Furthermore, CSI raw phase is ineffective for human sensing tasks due to the random phase
offset. Nevertheless, the phase information is indispensable for resolving the NLoS static problem because it is more sensitive than amplitude for a standstill person in the corner of the room [37]. For the above issues, the CSI ratio between transmission pairs is employed in our system as the SF. The CSI ratio matrix at the last timestamp is described as

$$\mathbf{F}_t^s = [r_t^1, r_t^2, \ldots, r_t^q, \ldots, r_t^Q],$$

where $q$ is the couple index, $Q \leq C_M \times N$ is the total number of the couples of transmission pairs, and $C$ is the combination calculator. Each CSI ratio vector in $\mathbf{F}_t^s$ is expressed as

$$r_t^q = [r_{t,q,1}, r_{t,q,2}, \ldots, r_{t,q,k}, \ldots, r_{t,q,K}].$$

According to the equation (3), each element of $r_t^q$ can be formulated as

$$r_{t,q,k} = \frac{h_{t,\alpha_1,k}}{h_{t,\alpha_2,k}} = \left| \frac{h_{t,\alpha_1,k}}{h_{t,\alpha_2,k}} \right| e^{j(\angle h_{t,\alpha_1,k} - \angle h_{t,\alpha_2,k})} = \frac{e^{-j\phi_{t,\alpha_1,k}} \sum_{l=1}^{L_{\alpha_1}} A_{t,\alpha_1,l} e^{-j2\pi d_{\alpha_1,l\alpha_1} \lambda_k}}{e^{-j\phi_{t,\alpha_2,k}} \sum_{l=1}^{L_{\alpha_2}} A_{t,\alpha_2,l} e^{-j2\pi d_{\alpha_2,l\alpha_2} \lambda_k}},$$

where $\alpha_1 = (m_1, n_1)$ and $\alpha_2 = (m_2, n_2)$ represent two transmission pairs, and $\alpha_1 \neq \alpha_2$. $m_1, m_2 \in [1, M]$ are the transmitting antenna indices, and $n_1, n_2 \in [1, N]$ are the receiving antenna indices. We can find out that the division of CSI complex values is equivalent to performing the division of amplitudes and the subtraction of phases. Afterward, since different antennas in COTS Wi-Fi devices share the same oscillator [38, 39], $r_{t,q,k}$ can be simplified as

$$r_{t,q,k} = \frac{\sum_{l=1}^{L_{\alpha_1}} A_{t,\alpha_1,l} e^{-j2\pi d_{\alpha_1,l\alpha_1} \lambda_k}}{\sum_{l=1}^{L_{\alpha_2}} A_{t,\alpha_2,l} e^{-j2\pi d_{\alpha_2,l\alpha_2} \lambda_k}}.$$

As shown in equation (13), the CSI ratio can eliminate the random phase offset term $e^{-j\phi_{t,\alpha_1,k}}$ and $e^{-j\phi_{t,\alpha_2,k}}$ in equation (12) because the random phase offsets for each transmission pair are the same, which makes the phase information available.

Fig. 1 shows the schematic diagram with the total number of antennas in the transmitter $M = 1$ and the receiver $N = 2$. The total number of propagation paths $L_{\alpha_1} = L_{\alpha_2} = 2$. The blue lines represent the LoS path which can receive the maximum power. On the other hand, the red lines are the NLoS path that reflects from the person to the receiver; while the dashed gray lines illustrate the NLoS path in an empty room. As can be observed in Fig. 1 a standstill person in NLoS scenarios leads to a change in CSI, including amplitude attenuation and phase shift. Hence, incorporating the amplitude and phase information is more useful when facing the NLoS static problem. CSI ratio removes the random phase offset and integrates amplitude and phase information when using the form of the complex value. Thus, the CSI ratio is more effective than amplitude
CSI ratio is a complex value and comprises the real and imaginary parts. To visualize the distinction between all cases, CSI ratio vector $r^t_q$ in equation (11) can be illustrated on a complex plane. One CSI vector has $K$ subcarrier points, and these $K$ points can constitute a shape on the complex plane. If all amplitude ratio and phase difference values of any two cases are different, these two CSI ratio vectors will form two unlike shapes in disparate areas. Hence, the pattern of CSI ratio on the complex plane can be generated as images for classification. In our system, the SF image generation function in Fig. 4 can generate the binary image $X^t_{B,q}$ for each CSI vector by mapping function $B(\cdot)$ as

$$X^t_{B,q} = B(r^t_q) : \mathbb{C}^K \rightarrow \mathbb{N}^{1 \times w \times h}, \quad (14)$$

where $w$ and $h$ mean the width and height of the image, respectively. $\mathbb{C}$ represents complex numbers. $X^t_{B,q}$ is a one-channel image, and subscript $B$ is denoted as the binary, indicating that each pixel of an image is either 0 or 1. Fig. 6 is a schematic diagram for understanding our designs. Data 1 and Data 2 denote two CSI ratio vectors with the number of subcarriers $K = 50$. Assume that owing to changes in the wireless channel, one CSI ratio vector forms a circle pattern, and the other is square. We can naturally know they are distinct by their shapes. Hence, as shown in the lower-left part of Fig. 6, the binary images containing the shape information can be generated to classify them. However, CSI ratio values can not be obtained in advance, so the SF image generation function intends to provide customized magnitude scalings and centroids for each image. Otherwise, the shape formed by CSI ratio points may be minified or cropped and even disappear when fixating the wrong
scalings and centroids, such as the erroneous images in Fig. 6. By the customized setting, the pattern in each graph is the principal part, and there are no more white areas increasing meaningless data. In conclusion, if human behaviors affect the shapes of the CSI ratios on the complex plane, binary images can be utilized as classification inputs.

b) Colorization: After generating the binary images, different cases can be discerned if their shapes are disparate. However, sometimes the patterns will be very similar, and the only distinction is their places on the complex plane. As depicted in the left part of Fig. 7, if Data 2 adds a slightly stronger reflection path relative to Data 1, all CSI ratio amplitudes of Data 2 have the same value increase, but their phases are almost unchanged. It will cause the CSI ratio of Data 2 to have the same shape as Data 1 but lead to a position offset to the upper right corner. This situation usually happens between the empty room and NLoS stationary case because the other two have large fluctuations that directly change their shapes. As shown in the lower-left part of Fig. 7, the binary images will not be able to deal with the problem that their patterns are identical. The reason is that the binary images only preserve the shape information, and the location information will be lost due to the specialized magnitude scalings and centroids when generating graphs. For this issue, the proposed colorization algorithm is designed to give location information to binary images by providing colors, further expanding the difference between the two cases of the empty room and NLoS stillness.

First of all, we utilize the addition of the real and imaginary parts as our location information because these two numbers represent the X-Y coordinates of a complex plane. The position value for each subcarrier in the CSI vector is defined as

$$p_{q,k}^t = Re\{r_{q,k}^t\} + Im\{r_{q,k}^t\}, \quad (15)$$

where $Re\{r_{q,k}^t\}$ and $Im\{r_{q,k}^t\}$ are the real and imaginary parts, respectively. As the name implies, the position value is the number entailing the location information of the complex plane. Before converting to colors, the range of position values mapped to the color bar needs to be determined first. If the bound is too wide or too narrow, all CSI ratio points will appear the same color on the complex plane. For this reason, our range of position values will be specified by the empty room case, and the time serial of the average CSI ratio in case 1 is expressed as

$$\bar{r}_q(e) = [\bar{r}_{q,1}(e), \bar{r}_{q,2}(e), \ldots, \bar{r}_{q,k}(e), \ldots, \bar{r}_{q,K}(e)]. \quad (16)$$

And each element of $\bar{r}_q(e)$ can be calculated as

$$\bar{r}_{q,k}(e) = \frac{1}{\tau_c} \sum_{\delta=0}^{\tau_c-1} r_{q,k}^\delta(e), \quad (17)$$

where $\tau_c$ is the window size for colorization, and $\delta$ is the sample index. This average computation is for acquiring the general trend of all subcarrier values during this period. After that, the maximum and minimum
position values are searched from the $\tilde{r}_q(e)$ in equation (16) with the rule of

$$p_{q,\text{max}} = \max_k (\text{Re}\{\tilde{r}_q(e)\} + \text{Im}\{\tilde{r}_q(e)\}),$$  \hspace{1cm} (18)$$

$$p_{q,\text{min}} = \min_k (\text{Re}\{\tilde{r}_q(e)\} + \text{Im}\{\tilde{r}_q(e)\}).$$  \hspace{1cm} (19)$$

max$_k$ and min$_k$ look for the maximum and minimum values along the subcarrier index $k$, respectively. In this way, most of the position values in case 1 do not exceed the range limited by $p_{q,\text{max}}$ and $p_{q,\text{min}}$. Eventually, we can transform the position value of each CSI ratio point into color, and the color matrix of the $q$-th couple can be presented as

$$C^t_q = [c^t_{q,1}, c^t_{q,2}, \ldots, c^t_{q,k}, \ldots, c^t_{q,K}].$$  \hspace{1cm} (20)$$

Each color in $C^t_q$ is formulated as

$$c^t_{q,k} = \begin{cases} 
  c_{q,\text{max}}, & p^t_{q,k} \geq p_{q,\text{max}} \\
  f(p^t_{q,k}), & p_{q,\text{min}} < p^t_{q,k} < p_{q,\text{max}} \\
  c_{q,\text{min}}, & p^t_{q,k} \leq p_{q,\text{min}} 
\end{cases},$$  \hspace{1cm} (21)$$

where $f(\cdot)$ is the mapping function from $R$ to $N^3$, and $R$ denotes the real number. The rainbow spectrum is adopted as our color bar, which indicates $c_{q,\text{max}} = f(p_{q,\text{max}})$ is red, and $c_{q,\text{min}} = f(p_{q,\text{min}})$ means purple. The three dimensions of $c^t_{q,k}$ represent the RGB channels of images. The benefits of colorization are displayed in the right part of Fig. 7. If the position value surpasses the defined upper or lower bound, it will correspond to red or purple. Since the range of the color bar is based on the empty room case, the images in case 1 such as Data 1 in Fig. 7 will possess a more uniform amount of colors. On the other hand, Data 2 in Fig. 7 has an offset, and the position values of some points exceed the boundary. We can observe that the image of Data 2 has more red areas. As a result, even though their shapes are identical, our colorization method can enlarge their visual difference and further distinguish them.

Finally, the RGB image of the $q$-th CSI ratio vector can be generated as

$$X^t_{\text{RGB},q} = D(X^t_{\text{B},q}, C^t_q): (N^{1\times w\times h}, N^{3\times K}) \rightarrow N^{3\times w\times h},$$  \hspace{1cm} (22)$$

where $D(\cdot)$ is the mapping function that can map the binary image and its colors into RGB images, and the three dimensions in $X^t_{\text{RGB},q}$ represent the RGB channel. The RGB images not only present the shape characteristics but also offer the location information through the color. Therefore, $X^t_{\text{RGB},q}$ is better than the binary image $X^t_{\text{B},q}$ when facing the NLoS static problem.

c) Channel Merging: Generally, using all couples data of the transmission pair to classify should have the best results. On the contrary, only taking a few couples for classification can achieve similar performance, and
a smaller amount of data is required. Moreover, the phase difference between some transmission pairs suffers a phase ambiguity which leads to a $\pi$ jump. The phase ambiguity results in a rotation of the CSI ratio pattern on the complex plane and will invalidate our design. Therefore, we only select some couples that do not suffer phase ambiguity so that our $Q \leq C_{2}^{M \times N}$. After getting the RGB images of all $Q$ couples, the classification input is the combination of these $Q$ pictures. Fig. 8 is a schematic diagram explaining how we conduct channel merging. If we directly concatenate $Q$ couples along the first dimension, the overall data volume will become very large because of the three channels of RGB images. This state will increase the computation time of our deep learning model. Hence, these RGB images will be converted into grey images with a channel number of 1 first. The grey images $X_{G,q}^{l}$ can be acquired by the mapping function $G(\cdot)$ as

$$X_{G,q}^{l} = G(X_{RGB,q}^{l}) : N^{3 \times w \times h} \rightarrow N^{1 \times w \times h}.$$ (23)

The reason why transforming the RGB to grey is that the grey image not only has a relatively small amount of data but also retains the effect of colorization. As the grey image shown in Fig. 8, each color converted to grayscale has its different shades, and we can find that red and purple in the grayscale will be the darkest color. When the CSI ratio vector has an offset, the darker area will become more. In this way, the grey images can still achieve the desired effect of our colorization. However, the binary image cannot be directly converted into a gray image at the beginning because the upper and lower bounds of the grayscale are white and black. If some cases appear in the white part, many points emerge white, and then they will not be visible. Thus, we still tend to generate RGB images when performing colorization and then convert them to greyscale in the
channel merging. Eventually, the grey images of all $Q$ couples can be merged as follows:

$$X^t_G = [X^t_{G,1}, X^t_{G,2}, \ldots, X^t_{G,q}, \ldots, X^t_{G,Q}]^T. \quad (24)$$

$X^t_G$ combines the CSI ratios of different couples to increase the richness of the data. Also, converting RGB images to gray images can reduce the quantity of data without sacrificing performance. As usual in image recognition, image normalization on $X^t_G$ is necessary. The input for our classification model can be expressed as $X^t_{\text{ratio}} = \frac{1}{s} X^t_G$, where $s$ is the maximum grayscale value for each pixel, and $t$ represents the last timestamp in the time window. In the following, we will refer to $X^t_{\text{ratio}}$ as the merged coloring CSI ratio image. Based on the above procedures, $X^t_{\text{ratio}}$ can be used to classify static cases because it includes the effect of colorization and the information between different couples.

### B. Three-Stage Supervised Contrastive Learning

After FEIG, the DF image $X_{rp}$ can be obtained to distinguish between static and dynamic cases, and the SF image $X^t_{\text{ratio}}$ can be generated to deal with the NLoS static problem. In this section, we will introduce our three-stage supervised contrastive learning in the blue block of Fig. 4, where inputs are $X_{rp}$ and $X^t_{\text{ratio}}$. Our supervised contrastive learning applies the framework in SupCon [25] and is a type of representation learning. In stage 1, we train the contrastive DF encoder that maps the RP into a representation vector according to different classes. This representation suggests the features learned from the RP. Subsequently, the contrastive SF encoder can be trained to map the merged coloring CSI ratio image to a representation vector in stage 2. Here, the consultation loss is designed to help further separate the representations of three static cases, which can improve the performance of our system. Finally, stage 3 utilizes the two representation vectors learned from stage 1 and stage 2 to complete classification. And the S3FEC in this stage can let our classifier automatically choose which representation to use in which case so that RPs and merged coloring CSI ratio images can perform their duties. The following is a detailed description of each stage.

1) **Stage 1 - Supervised Contrastive Learning**: In this stage, we focus on supervised contrastive learning for RPs. The core concept of contrastive learning is to bring the anchor and positive samples closer and push the anchor and negative samples farther in the embedding space. Thus, data with relatively similar features can be clustered into a group. And groups with different characteristics will be separated by a certain distance. Besides, the idea of 'supervised' is combining label information for contrastive learning. As described in SupCon, positives are defined as the samples with the same label as the anchor in a batch, while others with different labels are regarded as negative samples. As a result, SupCon can group the representations with the same label information in the embedding space, and the representation vectors between different labels will be pushed away from each other. Then, we can employ a simpler model to classify these trained representations.
Fig. 9 is a flow diagram of our network in stage 1. First, data augmentation is executed for the original RP to generate the augmented sample, which can be presented as

$$\tilde{X}_{rp} = \text{Aug}(X_{rp}) : N^{1 \times w \times h} \rightarrow N^{1 \times w \times h}. \quad (25)$$

Data augmentation can provide different views of the original image and increase the model generalization. Because RPs are binary images, our data augmentation operations only involve random resize cropping with the scale of $(\epsilon_1, \epsilon_2)$ and random horizontal flipping with the probability equaling $\varepsilon$ regardless of color. Note that our random resize cropping still will resize to the same size as the original image. After data augmentation, we can import the raw and augmented images into the same contrastive DF encoder. The representation of original sample can be formulated as

$$v_{rp} = \text{ResNet}^1(X_{rp}) : N^{1 \times w \times h} \rightarrow R^{512}, \quad (26)$$

where $\text{ResNet}^1$ is ResNet-18 [41] as the contrastive DF encoder in our work, and superscript 1 means stage 1. The representation vector of an augmented image can be expressed as $\tilde{v}_{rp} = \text{ResNet}^1(\tilde{X}_{rp})$. $v_{rp}$ and $\tilde{v}_{rp}$ in our work are vectors with 512 dimensions that can represent the feature of their images. Afterward, the
the contrastive SF encoder, and superscript 2 is described as stage 2. After that, their projection outputs are \( v_{rp} \) and \( 128 \). Notice that the operation in the left half is the same as stage 1. First, the augmented sample of original and augmented samples can be transformed as \( v_{rp} \) function also adopts random resize cropping and random horizontal flipping here. Next, the representations \( z_{rp} \) will make the data with the same label closer, and the data with different labels will become less similar. After the maximum value of 1, and the value of the denominator will become smaller. This result means that the loss their distance in the projection space. As loss decreases, the inner product in the numerator will be close to the other data. Computing the inner product can measure the similarity between two vectors, which also implies the normalization function that can let the vector lie on the unit hypersphere, which makes the maximum value of the inner product equal to 1. Also, the projection output of \( \tilde{v}_{rp} \) can be presented as \( \tilde{z}_{rp} = N(MLP^1(\tilde{v}_{rp})) \). Notice that \( v_{rp} \) is used instead of \( z_{rp} \) for classification in stage 3, but performing projection head is beneficial for contrastive loss [21]. Eventually, the supervised contrastive loss in stage 1 can be calculated based on \( z_{rp} \) and \( \tilde{z}_{rp} \). A dataset in a batch with a \( B \) batch size can be defined as \( \{ z_{rp,b}^1, y_{rp,b}^1 \}_{b=1,...,2B} \), where \( z_{rp,b}^1 \) is composed of original and augmented samples, and the sample index in this batch \( b \in [1, 2B] \) as a result. \( y_{rp,b}^1 \) indicates the label, and the label information of augmented data is the same as the original data. The superscript 1 in this dataset means it is in stage 1. Subsequently, the supervised contrastive loss can be formulated as \( L_{sc} = \sum_{i \in I} \frac{1}{P(i)} \sum_{p \in P(i)} \log \frac{\exp(z_{rp,i}^1 \cdot z_{rp,p}^1/\zeta)}{\sum_{a \in A(i)} \exp(z_{rp,i}^1 \cdot z_{rp,a}^1/\zeta)} \), where \( i \in I \equiv [1, 2B] \) is the anchor index, \( A(i) \equiv I \setminus i \) denotes the index set of other samples except the anchor, and \( P(i) \equiv \{ p \in A(i) | y_{rp,p}^1 = y_{rp,i}^1 \} \) is the index set of all positives. \( \tilde{P}(i) \) represents the cardinality of \( P(i) \), and \( \zeta \) is the scalar temperature parameter. As revealed in equation (28), the numerator and denominator respectively represent the inner product of anchor and positive samples and the inner product of anchor and other data. Computing the inner product can measure the similarity between two vectors, which also implies their distance in the projection space. As loss decreases, the inner product in the numerator will be close to the maximum value of 1, and the value of the denominator will become smaller. This result means that the loss will make the data with the same label closer, and the data with different labels will become less similar. After stage 1 training, the contrastive DF encoder can be learned to output the discriminative representations. And the well-trained \( v_{rp} \) will input the model in stage 3 to classify the four cases of human presence detection.

2) Stage 2 - Supervised Contrastive Learning and Consultation Loss: In stage 2, we focus on supervised contrastive learning for merged coloring CSI ratio images \( X_{ratio}^t \). Fig. 10 is the network flow diagram of stage 2. Notice that the operation in the left half is the same as stage 1. First, the augmented sample of \( X_{ratio}^t \) can be obtained by \( \tilde{X}_{ratio}^t = Aug(X_{ratio}^t) : R^{Q \times w \times h} \rightarrow R^{Q \times w \times h} \), and the data augmentation function also adopts random resize cropping and random horizontal flipping here. Next, the representations of original and augmented samples can be transformed as \( v_{ratio} = ResNet^2(X_{ratio}^t) : R^{Q \times w \times h} \rightarrow R^{512} \) and \( \tilde{v}_{ratio} = ResNet^2(\tilde{X}_{ratio}^t) \), respectively, where \( ResNet^2 \) is also ResNet-18 with input channel equaling \( Q \) as the contrastive SF encoder, and superscript 2 is described as stage 2. After that, their projection outputs are
expressed as \( z_{ratio} = N(MLP^2(v_{ratio})) \) and \( \tilde{z}_{ratio} = N(MLP^2(\tilde{v}_{ratio})) \), where the \( MLP^2 \) has the identical structure as \( MLP^1 \). Before computing the supervised contrastive loss, we also define the dataset of one batch with a \( B \) batch size in stage 2 as \( \{z_{ratio,b}^2, y_{ratio,b}^2\}_{b=1,...,2B} \). Again, \( z_{ratio,b}^2 \) contains raw and augmented samples. Finally, the supervised contrastive learning in equation (28) can be calculated by \( \tilde{z}_{ratio,b}^2 \) and \( y_{ratio,b}^2 \).

However, owing to the challenge of NLoS static problem, we need to add a supernumerary design to make the representations of empty and NLoS cases more separated in the embedding space. Therefore, the consultation loss is designed to assist supervised contrastive loss with training. The design concept of consultation loss comes from that distinguishing between walking and stationary situations is relatively simple in the problem of human presence detection. Our RPs are the appropriate answer to this problem, as the property depicted in Fig. 5. Consequently, the representations of the dynamic and static cases outputted from the contrastive DF encoder should be separated by distance after stage 1 training. This distance allows our model to classify these two categories easily, which can become a datum. Any two samples can be differentiated as long as the distance between their representations is close to this datum. According to this conception, if we can increase these two categories easily, which can become a datum. Any two samples can be differentiated as long as the distance between representation vectors of cases 1 and 2, the NLoS static problem may not be challenging. Later, since we intend to apply \( X'_{ratio} \) to classify three static situations, expanding the distances between three static cases is expected to be achieved here.

As shown in the right half of Fig. 10, the stage 1 encoder is necessary for stage 2 training, and both ResNet-18 and the projection head are essential. Their weights will not be updated in this stage because we need to calculate the well-trained distance between the dynamic and static RPs as a reference. Next, one batch dataset from the stage 1 encoder can be defined as \( \{z_{\text{rp},b}^c, y_{\text{ratio},b}^2\}_{b=1,...,B} \), where superscript \( c \) denotes consultation. This dataset does not contain augmented data because the reference distance is the only thing we need. The label \( y_{\text{ratio},b}^2 \) is the same as that in the dataset for CSI ratio, which corresponds to each other. Note that we calculate the consultation loss from \( z_{\text{rp}} \) and \( z_{\text{ratio}} \) can have less computation and leads to the same impact on representation because the projection head is just a dimension reduction function. In this way, the consultation loss can be formulated as

\[
L_{cs} = \frac{1}{S_s} \sum_{(i,j) \in S_s} d(z_{\text{ratio},i}^2, z_{\text{ratio},j}^2) - \frac{1}{S_{ds}} \sum_{(i,j) \in S_{ds}} d(z_{\text{rp},i}^c, z_{\text{rp},j}^c),
\]

where \( S_s(i,j) \equiv \{i,j \in [1,B]|i \neq j\} \) that satisfies \( (y_{\text{ratio},i}^2, y_{\text{ratio},j}^2) = (1,2) \lor (y_{\text{ratio},i}^2, y_{\text{ratio},j}^2) = (1,3) \lor (y_{\text{ratio},i}^2, y_{\text{ratio},j}^2) = (2,3) \) represents the indices set for the pairs of three static cases in a batch. The subscript \( s \) means the stationary, and the values of labels correspond to the four cases mentioned in section II-A. \( S_{ds}(i,j) \equiv \{i,j \in [1,B]|i \neq j\} \) is the indices set that meets \( (y_{\text{ratio},i}^2, y_{\text{ratio},j}^2) = (1,4) \lor (y_{\text{ratio},i}^2, y_{\text{ratio},j}^2) = (2,4) \lor (y_{\text{ratio},i}^2, y_{\text{ratio},j}^2) = (3,4) \). The subscript \( ds \) denotes dynamic and stationary, which indicates that one belongs to the moving case and the other belongs to the static situation in a batch. \( d(\cdot) \) is the function that
can calculate the Euclidean distance between two vectors, and \( \bar{S}_s \) and \( \bar{S}_{ds} \) are respectively the cardinalities of \( S_s(i, j) \) and \( S_{ds}(i, j) \). The left part of equation (29) is the average distance between the projection vectors \( z_{ratio} \) in three static situations, and the right part represents the average distance of \( z_{rp} \) between dynamic and static cases. Note that the stage 1 encoder will not be trained here. The distance calculated on the right half is usually greater than the distance on the left half. As the loss decreases, the absolute value of this subtraction will make the value of the left part closer to the right part. As a result, the distance between projection vectors from \( X_{ratio}^t \) of three static cases can be broadened, and we can acquire representations of the CSI ratio to classify these three cases. The process of this loss is like we refer to and consult the stage 1 encoder when training the contrastive SF encoder, so we named it consultation loss. Eventually, combined with supervised contrastive loss and consultation loss, the total loss of stage 2 can be expressed as

\[
L_{s2} = L_{sc} + \lambda L_{cs},
\]

where \( \lambda \) is the weight designed to control the influence of consultation loss. After stage 2 training, we can get better representations of merged coloring CSI ratio images than only employing supervised contrastive learning, which can be utilized for classification in stage 3 to resolve the NLoS static problem.

3) Stage 3 - Self-Switched Static Feature Enhanced Classifier (S3FEC): In the last stage, the thing we conduct is to classify the representations learned from stages 1 and 2. Generally speaking, after the representation learning, one representation vector can be classified efficiently by a simpler model \([21, 30]\), such as fully connected neural (FCN) networks and MLP. Nevertheless, our system contains two representations of \( X_{rp} \) and \( X_{ratio}^t \), which must involve the multi-view representation learning method. As described in \([42]\) and \([43]\), joint representation (JR) method that directly concatenates the representation vectors between different views is the most fundamental approach. Although this method is straightforward and has attained remarkable success, it cannot reach a consensus between different views. RPs are designed to be advantageous for distinguishing between dynamic and static cases, but the three stationary conditions are ambiguous for them. And merged coloring CSI ratio images are generated to classify between static situations, but it results in misclassification between moving and standing cases. The reason is that the CSI ratio is the data that contains only one timestamp. If the person walks to the corner or the LoS in the room, the signal characteristics of the CSI ratio will be very analogous to that in standing cases on the NLoS or LoS. Hence, directly concatenating these two representations for classification will cause them to restrict each other and can not improve performance. For this problem, the S3FEC is designed to let the model automatically use RPs when the person is moving and employ the coloring CSI ratio when differentiating three static cases.

Fig. 11 illustrates the stage 3 network flow diagram. The above two frozen ResNet-18 encoders trained from stages 1 and 2 are adopted to generate representations. Feature mapping can convert the representation vector into classification probability through one FCN and softmax activation function. Consequently, the classification
Fig. 11: Proposed Stage 3 Model.

probability of $X_{rp}$ can be expressed as

$$\hat{y}_d = [\hat{y}_{d,1}, \hat{y}_{d,2}, \hat{y}_{d,3}, \hat{y}_{d,4}]^T = \rho(v_{rp} \cdot W_{rp} + b_{rp}),$$

(31)

where subscript $d$ indicates the dynamic, and the four elements denote the predicted probabilities of four cases from RP. $\rho$ is the softmax activation function, and $W_{rp}$ and $b_{rp}$ are the weight and bias, respectively, in FCN for $v_{rp}$. And the classification probability of $X_{ratio}$ can be formulated as

$$\hat{y}_{ratio} = [\hat{y}_{ratio,1}, \hat{y}_{ratio,2}, \hat{y}_{ratio,3}, \hat{y}_{ratio,4}]^T = \rho(v_{ratio} \cdot W_{ratio} + b_{ratio}),$$

(32)

where $W_{ratio}$ and $b_{ratio}$ are the weight and bias of FCN for $v_{ratio}$. Each element in this vector is the output probability from the CSI ratio for each case. Here, we can acquire the two classification probabilities predicted by their representations, and the next step is crucial design in S3FEC. The model-learned switch can automatically exchange to let the appropriate probability become the final classification probability. The characteristics of RPs are very different in dynamic and stationary, so RPs will be more accurate in distinguishing these two categories. Ideally, feature mapping lets $\hat{y}_{d,4}$ be the largest element in the dynamic case while $\hat{y}_{d,4}$ is the smallest element in the static situations. In this way, the model can correctly judge whether the person is moving or
not. Therefore, as long as $\hat{y}_{d,4}$ is the maximum probability, the switch will make $\hat{y}_d$ the final classification probability. On the other hand, if $\hat{y}_{d,4}$ is not the maximum value, RPs consider this situation a static case. The switch will apply the probability generated by $X_{\text{ratio}}^r$ to classify. However, owing to the misclassification problem mentioned previously between moving and standing, directly taking $\hat{y}_{\text{ratio}}$ as the final probability may cause the result to be misjudged as case 4 again. To solve this issue, static feature enhancement is employed. The case 4 probability predicted by $v_{\text{ratio}}$ will be replaced with that outputted from $v_{\text{rp}}$. The reason is that RPs already perceive that no human is walking in the room, and the probability of case 4 will be a small value. When putting it into the dynamic element in $\hat{y}_{\text{ratio}}$, the final result can only be classified into static circumstances. As a result, the vector combining the dynamic probability from $\hat{y}_d$ and the stationary probability from $\hat{y}_{\text{ratio}}$ can be expressed as

$$\hat{y}_{\text{ratio}}' = [\hat{y}_{\text{ratio},1}, \hat{y}_{\text{ratio},2}, \hat{y}_{\text{ratio},3}, \hat{y}_{d,4}]^T.$$  

(33)

Since $\hat{y}_{\text{ratio}}'$ is not a probability vector, the softmax activation function will be employed again. The classification probability for static cases can be formulated as $\hat{y}_s = \rho(\hat{y}_{\text{ratio}}')$, where subscript $s$ indicates the static. In brief, the model-learned switch determines whether the dynamic probability from $\hat{y}_d$ is the maximum value. If so, the value of the switch is equal to 1, and let $\hat{y}_d$ be the final classification probability. Otherwise, the switch value is 0, and $\hat{y}_s$ becomes the final classification probability. Therefore, the final classification probability can be formulated as

$$\hat{y} = [\hat{y}_1, \hat{y}_2, \hat{y}_3, \hat{y}_4]^T = \omega \hat{y}_d + (1 - \omega) \hat{y}_s,$$  

(34)

where $\omega$ means the model-learned switch value. Eventually, the one-hot encoding output of the label can be expressed as $y = [y_1, y_2, y_3, y_4]^T$, and the cross-entropy loss for one batch can be calculated as

$$L_{ce} = - \sum_{i=1}^{B} \sum_{j=0}^{3} y_j(i) \log \hat{y}_j(i).$$  

(35)

As shown in Fig. 4, in addition to the well-trained contrastive DF and SF encoders, the online parameters of S3FEC in stage 3 can be obtained after offline training. In the online phase, after CSI data collection, real-time $X_{\text{rp}}$ and $X_{\text{ratio}}^r$ can be generated by FEIG described in section III-A. Subsequently, they can be imported into the online three-stage contrastive learning model. Finally, the final output $y_{\text{pre}}$ can be predicted from our system for human presence detection.

IV. PERFORMANCE EVALUATION

A. Experimental Setup

In our experiments, we evaluate the performance of our CRONOS system under two single-room scenarios in the actual environment. Two Wi-Fi routers with 2.447 GHz central frequency and 20 MHz bandwidth are
deployed in the indoor environment, one as the transmitter and the other as the receiver. Each router is equipped with two antennas. Thus, we can receive the CSI data of four transmission pairs in one packet, and the maximum number of couples for CSI ratios equals $C_2^4 = 6$. Moreover, each transmission pair contains $K = 56$ subcarrier values of the CSI data. With this setup, the data of four cases will be collected at the rate of 10 Hz, including the empty room, a NLoS standing person, a LoS standing person, and a walking person. Each class contains a few thousand data for training and testing, respectively.

The first experimental scenario will be carried out in a meeting room. As shown in Fig. 12, it is an indoor space with a size of $7 \times 6.75$ $m^2$, and we can observe that it is much larger than the room we live in our daily lives. The transmitter is placed on the podium in the upper right corner of the meeting room, while the receiver is put on the tripod in the lower left corner. This placement can make the two routers far away from
TABLE I: Training and testing data size for scene 1.

|            | Training | Testing |
|------------|----------|---------|
| Case 1     | 2000     | 2000    |
| Case 2 - upper left | 1000 | 500     |
| Case 2 - lower right | 1000 | 500     |
| Case 3     | 1000     | 500     |
| Case 4     | 2000     | 1000    |

TABLE II: Training and testing data size for scene 2.

|            | Training | Testing |
|------------|----------|---------|
| Case 1     | 1000     | 1000    |
| Case 2 - upper left | 500   | 500     |
| Case 2 - lower right | 500   | 500     |
| Case 4     | 1000     | 1000    |

The two NLoS corners and will let the NLoS static problem become severe. Next, the data sizes of four cases for training and testing are shown in Table I. The data in the empty case are collected when no one is in this meeting room, and the tester stands still in the corners of the upper left and the lower right orange blocks in Fig. 12(a) when gathering the data of NLoS static case. Furthermore, the pink block in Fig. 12(a) is where the tester stands still for collecting the LoS case, and the moving case is that the tester will walk around the green area. As shown in Fig. 13 the second experimental scenario is a normal-size office (4.7 \times 3.9 \ m^2), and the data sizes are also presented in Table II. Note that this office is relatively small, and the LoS stationary case is a rare situation in daily life. Thus, we will not collect the LoS static case in this scene.

TABLE III: System Parameters.

| Parameters                                           | Value |
|------------------------------------------------------|-------|
| Carrier frequency                                    | 2.447 \ GHz |
| Channel bandwidth                                    | 20 MHz |
| Total number of transmitter antennas \( M \)         | 2     |
| Total number of receiver antennas \( N \)           | 2     |
| Total number of subcarriers \( K \)                 | 56    |
| Data collection rate                                 | 10 Hz |
| Window size of RPs \( \tau \)                       | 50    |
| Window size for the threshold of RPs \( \tau_\gamma \) | [2000, 1000] |
| Threshold of RPs \( \gamma \)                       | [10, 33] |
| Total number of the couples of transmission pairs \( Q \) | [3, 2] |
| Size of image \( w \times h \)                       | 32 \times 32 |
| Window size for colorization \( \tau_c \)           | [2000, 1000] |
| Maximum value of each pixel \( s \)                 | 255   |
| Scale of random resize cropping \( (\epsilon_1, \epsilon_2) \) | (0.2, 1) |
| Probability of random horizontal flipping \( \epsilon \) | 0.5   |
| Batch size \( B \)                                  | 128   |
| Epoch for stages 1 and 2                             | 30    |
| Epoch for stage 3                                    | 10    |
| Scalar temperature parameter \( \zeta \)            | 0.07  |
| Weight for consultation loss \( \lambda \)          | 0.5   |

The parameters applied in our system are listed in Table III. The window size \( \tau \) of RPs is set as 50 to let
RPs contain enough timestamps, and the thresholds of RPs for scenes 1 and 2 are respectively set as 10 and 33 based on the CDF of 0.9 of the candidate set $D_e$ in equation (8). We will compare how different window sizes and thresholds affect the system performance in the following. Additionally, the window sizes $\tau_\gamma$ and $\tau_c$ in scene 1 are specified as 2000, and those in scene 2 are set as 1000. Note that their window sizes are the number of training data in the empty room of the two scenarios because we need to use all the training data to infer the general trend of the CSI signal. The number of the couples of transmission pairs $Q$ for CSI ratio is respectively set as 3 and 2 in these two scenarios. Therefore, it does not make the volume of input data too large and can avoid the $\pi$ jump in phase difference. Moreover, the size for RPs and coloring CSI ratio images is fixed to $32 \times 32$. Although the original size of the RP will be $50 \times 50$, we will resize it to $32 \times 32$ before inputting the model. In data augmentation, the scale of random resize cropping is set to (0.2, 1), and the probability of random horizontal flipping $\varepsilon$ is equal to 0.5. In the training process, the scalar temperature parameter $\zeta$ for supervised contrastive loss is set as 0.07 based on SupCon [25]. And the weight for the consultation loss $\lambda$ is selected as 0.5 to balance the impact of supervised contrastive loss and consultation loss on the model. According to the above parameter setting, we can achieve high accuracy in human presence detection.

### TABLE IV: The example of confusion matrix for each case.

| True   | Predicted | Positive | Negative |
|--------|-----------|----------|----------|
| Positive | True positives (TP) | False negatives (FN) |
| Negative | False positives (FP) | True negatives (TN) |

F1-score will be applied as our evaluation metric in the following experimental results. As shown in Table IV, true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) can be defined in the confusion matrix for each case. Next, we can get the recall $= \frac{TP}{TP+FN}$ and precision $= \frac{TP}{TP+FP}$, and the F1-score for each class can be calculated as

$$F1 - score = 2 \times \frac{recall \times precision}{recall + precision}.$$  

If we want to evaluate the overall performance of all cases, the average F1-score that averages the F1-scores of the four cases will be applied as the metric. It is worth mentioned that we will average the results of ten trials as the final result for each experiment. In the following, we firstly present the input images for each case and how some parameters affect our system. Then, we compare some different methods to show that the designs of our system are valid. Finally, we compare the recall and F1-score of the proposed CRONOS system to that of existing algorithm.
B. Input Image – RPs and Coloring CSI Ratio Images

As shown in Fig. 14, we present different input images of the four cases in scenario 1, including CSI amplitude image, binary CSI ratio image $X^t_{B,q}$, coloring CSI ratio image $X^t_{RGB,q}$, and RP $X_{rp}$. The CSI amplitude image is generated from a transmission pair, where the horizontal axis represents 56 subcarriers, and the vertical axis is their corresponding amplitude values. $X^t_{B,q}$, $X^t_{RGB,q}$, and $X_{rp}$ are generated by FEIG described in section III-A, where the CSI ratio only contains one couple. It is worth mentioned that RP is a graph that retains information about a time series, while the other three are pictures that only possess one timeslot information. First, we focus on NLoS static problem between the empty room and NLoS standing case. We can observe that the shapes in amplitude images are analogous between these two cases. The reason is that multipath signals reflected by the stationary person in the corner have a long propagation path and will be absorbed by the human body, which cannot cause much change to the CSI signal dominated by the LoS signal. Therefore, simply using amplitude is difficult to distinguish these two cases. Next, as the binary CSI ratio images of empty and NLoS cases shown in Fig. 14, they only have a slight difference in pattern, which indicates that the binary image that only retains the shape information cannot classify the NLoS situation from the empty case. However, because the colorization algorithm assigns the positional information of the complex plane to each point as a color, the coloring CSI ratio images have a detectable color change. The coloring image in empty case has an average area for each color, while the NLoS static case has more red and no purple points. Thus, coloring CSI ratio images for classification can resolve the NLoS static problem based on colors. On the other hand, since LoS halting or walking person will lead to a more significant influence
on the CSI signal, the CSI amplitude and ratio in these two cases will be different from that of the empty case. Hence, it is easy to separate them from the human absence case. Finally, we can see from RPs that the walking case is very distinguishable from the other three static cases. Because the walking situation will cause the amplitude difference of a time window to fluctuate, it will lead to more white areas in the RP. The CSI amplitude difference of the other three static cases is relatively stable, which makes their RPs almost black. Therefore, RPs can be employed to distinguish between dynamic and static cases. After merging all couples of CSI ratios by channel merging, we will input RPs and merged coloring CSI ratio images as shown in Fig. 14 into our three-stage contrastive model to predict the final human presence detection results.

C. Designed Parameters Evaluation

In our CRONOS system, some parameters may seriously affect the performance. Therefore, properly setting these parameters can make our system work normally and achieve high accuracy. In this section, we will analyze how these parameters affect our system from the data in scenario 1. According to the results, we can obtain the best parameters or verify that the parameter selection methods are valid for the CRONOS system. In the following analysis, the average F1-score of four cases is utilized as the metric for the parameter evaluation.

As depicted in Fig. 15, we compare the performance of different window size \( \tau \) in equation (6) and different threshold \( \gamma \) in equation (7) for RPs. The horizontal axis represents different threshold values, and the vertical axis is their F1-scores. The four lines with different shapes denote the results of four different window sizes. It is observed that the large or small threshold value leads to a performance reduction. The reason is that the threshold decides whether to make the point in RP white based on the amplitude difference. If \( \gamma \) is too small, the amplitude difference values will easily exceed the threshold. The RPs of all cases will have many white areas, which is impossible to distinguish dynamic and static situations with RPs. By contrast, if the threshold value is too large so that amplitude difference values in the moving case are always lower than it, the RPs of four classes are almost black. Therefore, we will set the threshold \( \gamma \) as 10 in scene 1 for the following
experiments, which makes the CDF $F_{D_c}(\gamma)$ equal to 0.9. It can let most of the RP points in static cases be black but let RP in the dynamic case has many white blocks, which can achieve the highest F1-score. Moreover, it can be seen that the smaller the window size, the worse the performance. As the curve with circle points illustrated in Fig. 15, $\tau = 5$ means that RPs only contain the information within 0.5 seconds due to the 10 Hz data collection rate. Because the effect of a walking human on the wireless channel is similar to the situation of a standing person during this short period, its performance is inferior. Generally speaking, the larger the window size, the more time information can be seen in one prediction. However, if it is too large, the response speed of our system will be reduced. Hence, $\tau = 50$ is the most feasible size in our system, which can have enough performance.

In this paragraph, we will compare the impact of different image sizes on performance and training time and observe how different $\lambda$ affect stage 2 training. In Fig. 16, the horizontal axis represents different image sizes. The left vertical axis shows the F1-score, and the right vertical axis means the training time. The five solid lines with different shapes are the results for different $\lambda$, while the dashed line represents the training time of five image sizes required for each epoch in stage 2. We can observe that the larger image size has a higher accuracy. The reason is that they have a higher resolution to present the features in the images. However, if the image size is too large, the training time for ResNet-18 will be very long. The time required for the $64 \times 64$ pixels images is nearly three times longer than that of $32 \times 32$ pixels, but their F1-scores are comparable. Thus, the size of $32 \times 32$ pixels is set as the most proper image size in the proposed CRONOS system. Next, it can be observed that if the weight for consultation loss is too large or too small, it will cause the system performance to decrease. A minuscule $\lambda$ means that the consultation loss is not helpful for the training of stage 2, which can not improve the performance. On the other hand, if $\lambda$ is considerable, the stage 2 training will be dominated by consultation loss. Without supervised contrastive loss, we are not able to separate the representations of different cases, resulting in performance degradation. Therefore, according to the results shown in Fig. 16, we will set the weight for consultation loss as 0.5, which can make the consultation loss effectively help training to get a better CSI ratio representation.

D. Effect of Different Input Images

In this section, we will compare the impact of different input images on the performance of our system. First, we will compare the performance of using only RPs and only merged coloring CSI ratio images and verify that our system needs these two images to help us attain high performance. Next, we will replace the $X^t_{ratio}$ with other input images to demonstrate that our CSI ratio with colorization algorithm can effectively solve the NLoS static problem and achieve higher human presence detection performance. In the following, these analyses will be performed on the data of scenario 1.
Fig. 17: The comparison between only $X_{rp}$, only $X_{ratio}^t$, and proposed CRONOS.

Fig. 18: The performance comparison between different input images.

Fig. 17 depicts the performance comparison between only inputting $X_{rp}$ and only using $X_{ratio}^t$ with the proposed CRONOS system. The horizontal axis represents four cases of human presence detection, and the vertical axis is their respective F1-score. Note that only using one image we designed will make the consultation loss and S3FEC impossible to implement. Hence, in these two single input cases, we only employ supervised contrastive loss to train their encoders and adopt the linear classifier with one FCN and softmax activation function for feature mapping. As shown in Fig. 17, we can observe that the results of three static cases are very deficient when only using $X_{rp}$, especially the LoS standing case. The reason is that the amplitude difference values in three static situations are relatively stable, and their values are not enough to exceed the threshold $\gamma$. Thus, as Fig. 14 depicts, their RPs will almost appear black, which leads to classification error. However, RP is very accurate when classifying the moving case and can achieve an F1-score of 97.83%. Hence, we can indeed apply RP to categorize the dynamic case from static ones. Furthermore, the F1-scores in three human presence situations are insufficient when only using $X_{ratio}^t$. As we mentioned in section III-B3, $X_{ratio}^t$ only includes the information of one packet, which gives rise to the ambiguity between human walking and standing and results in misclassification. Based on these results, it can be verified that our system requires $X_{rp}$ and $X_{ratio}^t$ to cooperate. Consultation loss and S3FEC can let them help each other and use their expertise to achieve higher performance.

Next, we will replace the $X_{ratio}^t$ in the CRONOS system with other graphs to compare their performance. As illustrated in Fig. 18, the substitute images include amplitude, phase, and binary CSI ratio images $X_{B,q}^t$ generated from equation (14). Notice that the amplitude and phase images are binary as the amplitude images presented in Fig. 14, and we will concatenate the results of the four transmission pairs to the input with four channels. For $X_{B,q}^t$, we also merge $Q$ couples to generate the merged binary CSI ratio images $X_B^t$ for evaluation. In this way, each type of image contains information about all transmission pairs. As can be seen in Fig. 18, the $X_B^t$ and CRONOS adopting the CSI ratio can provide higher F1-scores than amplitude and phase images when classifying empty and NLoS cases. The reason is that the CSI ratio eliminates the random phase
offset and combines the information of the amplitude and phase at the same time. Accordingly, the CSI ratio is more sensitive than pure amplitude or phase information for the stationary person in the corner. However, the performance of $X_B^l$ in LoS is worse than that of amplitude images. The explanation is that the person blocking the LoS will cause a more significant change in CSI amplitude than standing in NLoS. Hence, the erroneous classification usually occurs between empty and NLoS static cases when using amplitude images, and the LoS case is less likely to be misclassified. By comparison, the CSI ratio is more capable of detecting the NLoS standing person because it contains phase information. The difference between empty and NLoS cases becomes observable in the form of the complex plane image. Because the binary images only retain the difference in shape, the results will become misclassifications between empty, NLoS, and LoS cases, which leads to a lower F1-score in LoS standing case. Nevertheless, as long as the colorization algorithm provides the location information of the complex plane, the distinction between the three static cases can be magnified. We can get higher performance. In addition, it is observed that the F1-score of CRONOS is 19.39% higher than that of $X_B^l$ in the NLoS case, indicating that our colorization provides a beneficial effect for solving the NLoS static problem. Therefore, it is verified that the complex plane images of CSI ratio and colorization algorithm can let our system achieve better performance on human presence detection with NLoS static problem.

**E. Effect of Supervised Contrastive Loss and Consultation Loss**

Here, we will evaluate the benefits of supervised contrastive learning and consultation loss on the proposed CRONOS system. The data in scene 1 is also used for the following evaluation. First, we will compare the clustering effect of the representations learned with different loss combinations. Then, their performance will be provided to demonstrate that supervised contrastive loss can be better than the most commonly used cross-entropy loss on human presence detection. And the results show that the consultation loss can allow the contrastive SF encoder to learn better representations.

Firstly, we will provide visualizations of representations learned by different loss combinations, including cross-entropy loss for the entire three-stage contrastive model, the CRONOS system without consultation loss, and our proposed CRONOS system. The t-Distributed Stochastic Neighbor Embedding (t-SNE) [44] is applied to reduce the 512-dimensional representations into a 2D space for visualizations. As shown in Fig. [19], we will randomly sample 100 testing data for each case from the stage 2 contrastive SF encoder to display on the graph, where the four noticeable points denote the geometric centers of all testing data for each class. We can observe that only cross-entropy loss cannot separate the representations of the four cases. The justification is that cross-entropy loss only considers the classification results, and the learned representation vector will only give priority to the classification results. Thus, the data distribution of the four cases will be overlapped, which indicates that they are not efficacious representations. In addition, we also utilize the average Davies-Bouldin (DB) index [45] of 10 trials to evaluate the cluster validity of these three loss combinations. The smaller
Fig. 19: Representation comparison between different loss combinations: (a) only cross-entropy loss, (b) CRONOS system without consultation loss, and (c) the proposed CRONOS system.

Fig. 20: The performance comparison between different loss combinations.

DB index value is better because it indicates that the clustering results are close within the same cluster, and different clusters are far apart. The results reveal that our CRONOS system can achieve the lowest DB index, meaning that our consultation loss can effectively help us learn a better representation of the CSI ratio in stage 2. The CRONOS without consultation loss also has a lower DB index value than only using cross-entropy loss, implying that the supervised contrastive loss can push the representation of different classes further to learn the representation with a better classification effect. Finally, we also calculate the Euclidean distance between the geometric centers of the three static cases in the experiment. The Euclidean distances of CRONOS between cases 1 and 2, cases 1 and 3, and cases 2 and 3 are 68.95, 95.79, and 97.4, respectively, while the Euclidean distances without consultation loss are respectively 65.87, 80.24, and 82.85. According to this result, it can also be shown that our consultation loss can increase the distance between the representations of the three statics. More importantly, the distance between the empty room case and NLoS static situation can be increased, which can help resolve the NLoS static problem on human presence detection.
After observing their respective representations, we can compare their performance. As illustrated in Fig. 20, it can be seen that the overall performance of both methods with supervised contrastive learning is higher than that with only cross-entropy loss. The reason is that if only cross-entropy is used to train a model that combines contrastive DF and SF encoders, the learning model extract features only from the input images. Consequently, as shown in Fig. 19(a), only using cross-entropy loss will learn an inadequate representation. By contrast, supervised contrastive learning first separates the four cases by contrastive loss and obtains better representations. Then, using cross-entropy loss to classify these representations can attain better classification results. Moreover, consultation loss can improve the F1-score by 2.1%, 6.47%, and 0.06% in cases 1, 2, and 3, respectively. The rationale is that the consultation loss refers to the Euclidean distances between dynamic and static classes from the RPs when training the stage 2 encoder, thereby increasing the distance between the representations of the three static cases. This result also verifies that the consultation loss can provide some assistance in solving the NLoS static problem. Therefore, the CRONOS system with supervised contrastive loss and consultation loss can achieve the highest performance on human presence detection.

F. Effect of Different Model Architectures

In CRONOS, we employ ResNet-18 as our contrastive DF and SF encoders and S3FEC as the classifier. In this section, we will first use the testing data of scenario 1 to examine whether our model-learned switch performs according to our arguments. Then, different model architectures are compared to verify that the designed architectures of our system are necessary and beneficial.

As displayed in Fig. 21, we illustrate the switch values without S3FEC and with S3FEC, where the horizontal axis represents the index of testing data in scenario 1, and the vertical axis is the switch values. It is worth mentioned that the switch value $\omega$ is either 0 or 1. In S3FEC, $\omega = 1$ means that the model uses the $\hat{y}_d$ as the
Fig. 22: The performance comparison between different model architectures.

final classification probability, while $\omega = 0$ denotes that the $\hat{y}_s$ is utilized. On the other hand, the switch values without S3FEC come from an additional FCN, where its input is the concatenation of representations $v_{rp}$ and $v_{ratio}$. Under this method, the switch value of 1 also means that the model chooses $\hat{y}_d$ as the final classification probability, while the switch value of 0 implies that the model applies the $\hat{y}_{ratio}$ from equation (32). In Fig. 21(a), it is observed that the model without S3FEC will employ more probabilities from RPs to classify the three static cases. We know that the RPs of the three static cases are very similar so that three static cases will be misclassified from each other. When classifying some data in the moving situation, the probability from the CSI ratio is used as the final probability. As presented in Figure 4.6, applying the probabilities generated by the CSI ratio may result in the misclassification of three human presence situations. Therefore, the method without S3FEC has a lower F1-Score of 82.04%. By comparison, the switch values in S3FEC can be implemented according to our designs. As shown in Fig. 21(b), the probability from the CSI ratio is used to classify three static cases, and that generated by RPs is chosen when the person is walking. Therefore, the proposed CRONOS system with S3FEC can acquire a higher F1-score of 95.24% since S3FEC allows $X_{rp}$ and $X_{ratio}^t$ to classify the cases they are capable of handling.

Next, we will evaluate how different model architectures would affect our system. First, we will replace the ResNet-18 of stages 1 and 2 with a two-layer CNN architecture with a kernel size of 5. Second, S3FEC will be superseded by the JR method that concatenates two representation vectors with a linear classifier. For the last one, we will implement an attention-based approach. This method multiplies $\hat{y}_d$ and $\hat{y}_{ratio}$ by two weight vectors that come from inputting their respective representations into an FCN. As depicted in Fig. 22, the two-layer CNN encoders will lead to an inoperable result. The reason is that their losses cannot be decreased in stages 1 and 2, which indicates that the two-layer CNN architecture is too simple to perform more complex training such as supervised contrastive learning. Hence, a more complex ResNet is necessary for our system. In addition, we can observe that the performance of the JR method is lower than that of the proposed CRONOS system except the empty case. This explanation is that RPs are only good at classifying dynamic and static classes, and merged coloring CSI ratio images are good at distinguishing three static situations. As
described in section III-B3, directly concatenating their representations will restrict each other. It will result in misclassification between the three static cases, and the performance of the moving class will not be better than our CRONOS system. By contrast, the proposed S3FEC lets RRs determine whether someone is walking and lets the CSI ratio classify the remaining three static cases. They don’t limit each other and allow them to do their suitable parts; thus, S3FEC outperforms the JR method. On the other hand, the attention-based method is less effective. It was hoped that the learned weight vectors would allow the model to implement the function of S3FEC. However, the result reveals that the model cannot obtain the weights with the function of S3FEC during training, which in turn makes the classification performance lower. For this reason, our system needs to add the model-learned switch and static feature enhancement mechanism in S3FEC to achieve the highest performance.

G. Overall Performance Comparison

In the last section, we will compare the overall performance of our proposed CRONOS with some baselines in two experimental scenes. LC-DNN [46] generates a position-dependent local feature (PDL-feature) based on CSI amplitude for indoor localization. A local connection based deep neural network (LC-DNN) is designed to extract the features of adjacent subcarriers to improve performance. We will employ the same input and model architecture, but the classification result will become the four cases of human presence detection. F-LSTM [33] first utilizes a Butterworth low-pass filter on the CSI amplitude to suppress noise, then uses a moving average filter to reduce the influence of large amplitude or burst noises. After that, the LSTM model is employed to classify these filtered CSI amplitudes for human presence detection. Next, P-CNN [34] first performs some data preprocessing on CSI amplitude and phase information and converts them into images, and two parallel CNNs are employed for these two images to implement human presence detection. Moreover, C-MuRP [36] processes the CSI amplitude and inputs it into a conditional gated recurrent unit (GRU) for multi-room indoor person detection. Finally, CALPD [47] is our previous work, using coloring CSI ratio images and a CNN model for NLoS human presence detection. The following is the performance comparison on recall and F1-score between the proposed CRONOS and baselines.

As shown in Table V, we compare the recall and F1-score of our proposed CRONOS with other existing methods on the dataset in scenario 1. Notice that the F1-score is abbreviated as F1 in Table V and the $\Delta$F1 represents the difference with the F1-score of CRONOS. We additionally present the F1-scores of the NLoS case and average result as a bar chart in Fig. 23. In Table V we can observe that CRONOS outperforms the baselines in classifying empty and NLoS static cases. Specifically, the proposed CRONOS system can improve 0.53% recall and 0.27% F1-score compared with the best baseline C-MuRP in the human absence case. More importantly, our CRONOS yields a 19.85% improvement in recall and 18.63% in F1-score compared with the best baselines CALPD and C-MuRP, respectively. This result also demonstrates that our system is more capable
TABLE V: The overall performance comparison in scenario 1.

| Approach  | Empty Room | NLoS standing case | LoS standing case | Moving case | Average |
|-----------|------------|---------------------|-------------------|-------------|---------|
|           | Recall     | F1                  | ∆F1               | Recall      | F1      | ∆F1     | Recall   | F1      | ∆F1     | Recall   | F1      | ∆F1     |
| LCDNN     | 91.38      | 83.6                | 12.19             | 48.54       | 56.29   | 33.47   | 84.14    | 84.29   | 12.59   | 80.12    | 79.33   | 19.36   | 76.04    | 75.88   | 19.4    |
| F-LSTM    | 90.74      | 85.31               | 10.48             | 44.45       | 50.55   | 39.21   | 88.23    | 89.43   | 7.45    | 91.01    | 89.35   | 9.34    | 78.61    | 78.66   | 16.2    |
| P-CNN     | 49.56      | 55.77               | 40.02             | 38.51       | 26.62   | 63.14   | 35.26    | 51.81   | 45.07   | 99.98    | 99.99   | -1.3    | 55.83    | 58.55   | 36.73   |
| C-MuRP    | 97.43      | 95.52               | 0.27              | 57.56       | 71.13   | 18.63   | 96.56    | 97.5    | -0.62   | 97.75    | 84.34   | 14.35   | 87.32    | 87.12   | 8.16    |
| CALPD     | 80.7       | 82.16               | 13.63             | 66.08       | 65.55   | 24.21   | 77.08    | 80.55   | 16.33   | 84.92    | 79.48   | 19.21   | 77.19    | 76.94   | 18.34   |
| w/o Xratio| 92.14      | 84.8                | 10.99             | 78.14       | 72.12   | 17.64   | 0        | 0       | 96.88   | 95.99    | 97.83   | 0.86    | 66.57    | 63.69   | 31.59   |
| w/o Xrp   | 95.44      | 95.14               | 0.65              | 42.41       | 56.63   | 33.13   | 50.71    | 66.21   | 30.67   | 91.62    | 68.15   | 30.54   | 70.04    | 71.53   | 23.75   |
| w/o Lsc   | 96.6       | 92.76               | 3.03              | 71.18       | 75.31   | 14.45   | 99.93    | 96.17   | 0.71    | 98.29    | 98.81   | -0.12   | 91.5     | 90.76   | 4.52    |
| w/o Lcs   | 98.15      | 93.69               | 2.1               | 75.8        | 83.29   | 6.47    | 94.9     | 96.82   | 0.06    | 97.38    | 97.49   | 1.2     | 91.56    | 92.82   | 2.46    |
| w/o S3FEC | 98.9       | 95.66               | 0.13              | 81.43       | 86.37   | 3.39    | 70.98    | 81.84   | 15.04   | 98.88    | 94.14   | 4.55    | 87.55    | 89.5    | 5.78    |

CRONOS: 97.96  95.79  -  85.93  89.76  -  96.79  96.88  -  98.16  98.69  -  94.71  95.28  -

Fig. 23: The performance comparison with baselines on the NLoS case and average result in scenario 1.

of resolving the NLoS static problem than other baselines. Although CRONOS is not the most efficient in LoS static and moving cases, the recall and F1-score of our system are nearly the same as the best baseline. Therefore, as depicted in Fig. 23, our CRONOS system outperforms all baseline methods on average performance. This result indicates that our system can not only deal with the NLoS static problem but also classify other human presence detection cases. As presented in the lower half of the Table V, we also conducted ablation studies to verify the validity of our designs. It can be seen that the lack of any one strategy will lead to a drop in F1-score except for the walking case. For example, removing $X_{rp}$ or $X_{ratio}$ will seriously degrade performance. As we mentioned earlier, this reason is that they have their unclassifiable classes. In addition, removing supervised contrastive learning leads to a decrease in overall performance. It indicates that representations trained first with supervised contrastive loss can more effectively improve classification performance. Moreover, removing the consultation loss will decrease the F1-scores of three static cases, especially the 6.47% decline in the NLoS standing case. This result implies that our consultation loss can increase the Euclidean distances of representations between three static classes, which makes their classification results better. Finally, the ablation study on S3FEC also demonstrates that S3FEC can allow RPs and CSI ratio to perform their duties. Therefore, CRONOS adopting all designs can have the highest 94.71% recall and 95.28% F1-score on human presence detection.
TABLE VI: The overall performance comparison in scenario 2.

| Approach      | Empty Room | NLoS standing case | Moving case | Average |
|---------------|------------|--------------------|-------------|--------|
|               | Recall     | F1   | ∆F1               | Recall     | F1   | ∆F1              | Recall     | F1   | ∆F1              |
| LC-DNN [46]   | 77.64      | 77.9 | 20.86             | 43.02      | 47.69 | 50.54            | 52.98      | 46.64 | 52.83            |
| F-LSTM [33]   | 96.91      | 91.91 | 8.65              | 34.17      | 47.21 | 51.02            | 90.41      | 82.04 | 17.43            |
| P-CNN [34]    | 92.89      | 90.47 | 8.29              | 85.43      | 88.17 | 10.06            | 100        | 100   | -0.53            |
| C-MuRP [36]   | 97.51      | 91.77 | 6.99              | 25.05      | 39   | 59.23            | 97.35      | 74.62 | 24.85            |
| CALPD [47]    | 95.72      | 93.1 | 5.66              | 89.54      | 89.87 | 8.36             | 82.44      | 84.49 | 14.98            |
| w/o $X_{t_{\text{st}}}^{t}$ | 67.29 | 61.99 | 36.77            | 47.57      | 50.58 | 47.65            | 99.56      | 99.66 | -0.19            |
| w/o $X_{r_{p}}$ | 94 | 93.34 | 5.42              | 65.31      | 73.73 | 24.5             | 93.57      | 82.74 | 16.73            |
| w/o $L_{sc}$  | 99.12      | 96.48 | 2.28              | 89.52      | 93.68 | 4.55             | 100        | 98.29 | 1.18             |
| w/o $L_{cs}$  | 99.82      | 98.23 | 0.53              | 95.63      | 97.51 | 0.72             | 99.61      | 99.4 | 0.07             |
| w/o S3FEC     | 99.81      | 98.57 | 0.19              | 94.8       | 96.89 | 1.34             | 99.04      | 98.26 | 1.21             |
| CRONOS        | 99.79      | 98.76 | -                 | 97.17      | 98.23 | -                | 99.44      | 99.47 | -                |

Fig. 24: The performance comparison with baselines on the NLoS case and average result in scenario 2.

As shown in Fig. 13, scenario 2 is a smaller indoor space, and we additionally perform performance evaluation in this scene. Note that we did not evaluate the LoS static case in this scene since the probability of a human standing still in LoS is relatively small in this room. In Table VII and Fig. 24, the proposed CRONOS still outperforms all baselines in NLoS case and average performance. In the NLoS standing case, the recall and F1-score of our system are respectively 7.64% and 8.36% higher than the best baseline CALPD. This result reveals that our system can still effectively solve the NLoS static problem in this scenario. Furthermore, it can be seen that our CRONOS outperforms the P-CNN by 6.03% and 5.94% in terms of average recall and F1-score, respectively. Eventually, the ablation study in this scenario also has similar results to scenario 1, which verifies that all designs have their benefits in this scenario. Therefore, even though the experimental scene is changed to a smaller room, our system with the designed parameter setting can still achieve the highest performance. The experimental results also verify the feasibility of our algorithm for human presence detection.

V. Conclusion

In this thesis, we propose a device-free human presence detection system using Wi-Fi CSI signals, which can detect a human standing still in NLoS corners. Specifically, our proposed CRONOS consists of FEIG and
three-stage supervised contrastive learning. FEIG first generates RPs from CSI amplitude difference to help us distinguish dynamic from static situations. Afterward, CSI ratio complex plane images are generated to classify the three static cases, and colorization is proposed to enhance the visual difference between empty and NLoS classes to resolve NLoS static problem. Moreover, supervised contrastive loss is utilized in three-stage supervised contrastive learning to train the representations of RPs and merged coloring CSI ratio images. The learning with consultation loss can expand the distance between the representations of the three static cases to improve performance. Finally, S3FEC employs the model-learned switch and static feature enhancement mechanism to allow the model to automatically select the probability from RP or merged coloring CSI ratio image as the final classification probability. Experimental evaluations demonstrate that our system can achieve the highest average F1-score of 95.28%. The 18.63% improvement compared with the best baseline in the NLoS case indicates that the proposed CRONOS system can effectively implement human presence detection for the NLoS static person. In the future, we plan to convert supervised learning to self-supervised, which can help us reduce the labor of data collection. In addition, additional data augmentation designs are expected to employ for human presence detection. Finally, we aim to extend the system to multi-room and multi-person situations, hoping that a pair of Wi-Fi routers can detect NLoS standing people in two rooms at the same time.

REFERENCES

[1] Y. Wang, S. Yang, F. Li et al., “FallViewer: A Fine-Grained Indoor Fall Detection System with Ubiquitous Wi-Fi Devices,” IEEE Internet of Things Journal, vol. 8, no. 15, pp. 12455–12466, Mar. 2021.

[2] Z. Tian, Y. Li, M. Zhou et al., “WiFi-based Adaptive Indoor Passive Intrusion Detection,” in Proceedings IEEE International Conference on Digital Signal Processing (ICDSP), Nov. 2018, pp. 1–5.

[3] Z. Zhiqiang, G. Xuebin, B. Jit et al., “Moving Targets Detection and Localization in Passive Infrared Sensor Networks,” in Proceedings IEEE International Conference on Information Fusion (FUSION), Dec. 2007, pp. 1–6.

[4] V. G. Moshnyaga, K. Hashimoto, and T. Suetsumi, “A Hardware Design of Camera-based User’s Presence Detector,” in Proceedings IEEE International Conference on Systems, Man and Cybernetics (SMC), Apr. 2008, pp. 429–432.

[5] J. Xiao, K. Wu, Y. Yi et al., “FIFS: Fine-Grained Indoor Fingerprinting System,” in Proceedings IEEE International Conference on Computer Communications and Networks (ICCCN), Aug. 2012, pp. 1–7.

[6] R. Zhou, M. Hao, X. Lu et al., “Device-Free Localization based on CSI Fingerprints and Deep Neural Networks,” in Proceedings IEEE International Conference on Sensing, Communication, and Networking (SECON), Jun. 2018, pp. 1–9.

[7] X. Wang, L. Gao, and S. Mao, “CSI Phase Fingerprinting for Indoor Localization with a Deep Learning Approach,” IEEE Internet of Things Journal, vol. 3, no. 6, pp. 1113–1123, Apr. 2016.

[8] J. Xi, Z. Xu, L. Chen et al., “Human Counting and Action Recognition with WiFi via Deep Learning,” in Proceedings IEEE Cross Strait Radio Science and Wireless Technology Conference (CSRSWTC), Mar. 2021, pp. 1–3.

[9] Y. Yang, J. Cao, X. Liu et al., “Wi-Count: Passing People Counting with COTS WiFi Devices,” in Proceedings IEEE International Conference on Computer Communication and Networks (ICCCN), Oct. 2018, pp. 1–9.

[10] J. Liu, Y. Chen, Y. Wang et al., “Monitoring Vital Signs and Postures During Sleep using WiFi Signals,” IEEE Internet of Things Journal, vol. 5, no. 3, pp. 2071–2084, Apr. 2018.

[11] X. Wang, C. Yang, and S. Mao, “PhaseBeat: Exploiting CSI Phase Data for Vital Sign Monitoring with Commodity WiFi Devices,” in Proceedings IEEE International Conference on Distributed Computing Systems (ICDCS), Jul. 2017, pp. 1230–1239.
[35] Y.-M. Huang, A.-H. Hsiao, C.-J. Chiu et al., “Device-Free Multiple Presence Detection using CSI with Machine Learning Methods,” in Proceedings IEEE Vehicular Technology Conference (VTC2019-Fall), Nov. 2019, pp. 1–5.

[36] F.-Y. Chu, C.-J. Chiu, A.-H. Hsiao et al., “WiFi CSI-based Device-Free Multi-Room Presence Detection using Conditional Recurrent Network,” in Proceedings IEEE Vehicular Technology Conference (VTC2021-Spring), Jun. 2021, pp. 1–5.

[37] H. Wang, D. Zhang, Y. Wang et al., “RT-Fall: A Real-Time and Contactless Fall Detection System with Commodity WiFi Devices,” IEEE Transactions on Mobile Computing, vol. 16, no. 2, pp. 511–526, Apr. 2017.

[38] N. Tadayon, M. T. Rahman, S. Han et al., “Decimeter Ranging with Channel State Information,” IEEE Transactions on Wireless Communications, vol. 18, no. 7, pp. 3453–3468, May. 2019.

[39] M. Kotaru, K. Joshi, D. Bharadia et al., “SpotFi: Decimeter Level Localization using WiFi,” in Proceedings ACM Conference on Special Interest Group on Data Communication (SIGCOMM), Aug. 2015, p. 269–282.

[40] H. Zhu, Y. Zhuo, Q. Liu et al., “π-Splicer: Perceiving Accurate CSI Phases with Commodity WiFi Devices,” IEEE Transactions on Mobile Computing, vol. 17, no. 9, pp. 2155–2165, Jan. 2018.

[41] K. He, X. Zhang, S. Ren et al., “Deep Residual Learning for Image Recognition,” arXiv:1512.03385, Dec. 2015.

[42] Y. Li, M. Yang, and Z. Zhang, “A Survey of Multi-View Representation Learning,” IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 10, pp. 1863–1883, Sep. 2019.

[43] X. Jia, X.-Y. Jing, X. Zhu et al., “Semi-Supervised Multi-View Deep Discriminant Representation Learning,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 7, pp. 2496–2509, Feb. 2021.

[44] L. Van der Maaten and G. Hinton, “Visualizing Data using t-SNE,” Journal of Machine Learning Research, vol. 9, no. 86, pp. 2579–2605, Jun. 2018.

[45] D. L. Davies and D. W. Bouldin, “A Cluster Separation Measure,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.

[46] W. Liu, H. Chen, Z. Deng et al., “LC-DNN: Local Connection based Deep Neural Network for Indoor Localization with CSI,” IEEE Access, vol. 8, pp. 108720–108730, Jun. 2020.

[47] C.-C. Hsieh, A.-H. Hsiao, C.-J. Chiu et al., “CSI Ratio with Coloring-Assisted Learning for NLoS Motionless Human Presence Detection,” in Proceedings IEEE Vehicular Technology Conference (VTC2022-Spring), Aug. 2022, p. 1–5.