SAPIENZA
UNIVERSITÀ DI ROMA

Person Re-ID through Radio Biometric Signatures, Human Silhouette and Skeleton Video Synthesis through Wi-Fi Signals.

Faculty of Information Engineering, Informatics, and Statistics

Doctor of Philosophy in Computer Science – XXXIV Cycle

Candidate

Marco Cascio
ID number 1715706

Thesis Advisors

Prof. Luigi Cinque
Prof. Chiara Petrioli

Co-Advisor

Prof. Danilo Avola

Person Re-ID through Radio Biometric Signatures, Human Silhouette and Skeleton Video Synthesis through Wi-Fi Signals.

Ph.D. thesis. Sapienza – University of Rome

Dedicated to
my family, the March family, Ariel, and Danilo Avola.

Abstract

The increasing availability of wireless access points (APs) is leading to human sensing applications based on Wi-Fi signals as support or alternative tools to the widespread visual sensors, where the radio signals enable to address well-known vision-related problems, such as illumination changes or occlusions, and human privacy concerns. These are possible because both objects and people affect wireless signals, causing radio variations observed through the channel state information (CSI) measurement of Wi-Fi APs that allows signal-based feature extraction, e.g., amplitude or phase. On this account, this thesis shows how the pervasive Wi-Fi technology can be directly exploited for solving person Re-ID, for the first time in literature, and image synthesis problems. More accurately, for the former, Wi-Fi signals amplitude and phase are extracted from CSI measurements and analyzed through a two-branch deep neural network working in a siamese-like fashion. The designed pipeline can extract meaningful features from signals, i.e., radio biometric signatures, that ultimately allow the person Re-ID. The effectiveness of the proposed system is evaluated on a specifically collected dataset, where remarkable performances are obtained; suggesting that Wi-Fi signal variations differ between different people and can consequently be used for their re-identification. Instead, for the image synthesis, a novel two-branch generative neural network effectively maps radio data into visual features, following a teacher-student design that exploits cross-modality supervision. This strategy conditions signal-based features in the visual domain to completely replace visual data. Once trained, the proposed method synthesizes human silhouette and skeleton videos using exclusively Wi-Fi signal amplitude. The approach is evaluated on publicly available data, where it obtains remarkable results for both silhouette and skeleton videos generation, demonstrating the effectiveness of the proposed cross-modality supervision strategy. Concluding, a small use case is specified to show how videos synthesized from wireless signals can be used to solve human activity recognition obtaining a privacy-conscious system and potentially exploiting the radio signal benefits.

Contents

Chapter 1

Introduction

In recent years, Wi-Fi technology is becoming ever more present equally in private and public places by promoting, other than the constant Internet connection, the growth of the Internet of Things (IoT) and wireless sensing applications. Wi-Fi sensing is new automation enabling perception- and understanding-based tasks by exploiting the now ubiquitousness of such wireless signals. The latter gained momentum as an alternative or support solution to the classical human-related comprehension problems traditionally based on visual information [8], such as event detection [188, 177], gesture [182, 179] and activity recognition [143, 41], localization [136, 81] and tracking [157, 121], health-care [173, 2], or pose estimation [65, 1]. The reason is twofold. First, human privacy is naturally preserved because sensitive information is not collected through the radio-based medium. Second, wireless technology is not affected by eventually challenging visibility conditions. Generally, vision-based applications are constrained to the frame of low-resolution or costly visual sensors, requiring multiple devices strategically located in different places to cover different points of view, even for small limited areas. In contrast, wireless sensing applications can cover vast areas, potentially exploiting the existing networks built through commodity, commercial, and low-cost Wi-Fi devices. Indeed, even the IEEE 802.11 wireless LAN working group is already focusing on WLAN sensing applications [104], and in the near future, Wi-Fi might become a day-life sensing technology. Among the others, two of the most exciting fields still never or barely explored through Wi-Fi sensing are person re-identification and image synthesis [67, 50].

Person re-identification (Re-ID) addresses a recognition task across non-overlapping camera views, understanding whether a given person appeared in the same (or a different) location at distinct time instants [24]. Direct evolution from identification approaches, where a person identity is classified into one of those known by the system, in a Re-ID method the input image, called probe, is matched against a gallery of identities so that the most probable one can be retrieved, as can be observed in Fig. 1.1. While difficult, this task naturally enables for the re-identification of people that were never seen before. To achieve this goal, however, it is necessary to correctly model a person's appearance and, as a consequence, existing approaches exploit distinctive and reliable visual features extracted from images and video sequences [72, 169]. Although these features have achieved impressive results in Re-ID over the last decade, especially due to deep learning advances, several challenges are still open, including different viewing angles [33, 123]; illumination changes [145, 34]; background clutter [116, 191]; occlusions [91, 58]; and long-term Re-ID [77, 60], where the person's appearance can drastically change after long periods of time (e.g., weeks). What is more, even though ever improving methodologies are being developed to address these challenging problems [172, 94, 75], person Re-ID is still considered an open task and, even more aggravating, it also presents a considerable gap between research-oriented and practical scenarios [72]. Nevertheless, great efforts are being made to improve this situation by also investigating techniques based on different prerequisites. Indeed, alternative solutions are already exploiting skeleton information [13] or multiple and diverse technologies such as thermal and infrared images [168, 170] or radar sensors [31], since the re-identification task can be a crucial asset in real application areas such as surveillance and forensics. To expand on this matter, can be explored an unorthodox technology that, due to its nature, inherently avoids the vision-based complications, and introduce a novel approach based on a different medium, i.e., Wi-Fi transmissions.

Once the person is re-identified, perceiving and understanding the human behavior can be an immediate consequence in advanced people monitoring scenarios. Therefore, such radio transmissions can potentially be used even to synthesize human-related images later employed for analyzing the human activities. Generally, the image synthesis task refers to the creation of a new image using different types of sources as an image description, as shown in Fig. 1.2. In recent years, the state-of-the-art proposed several promising deep learning methods, including, but not limited to, text-to-image [178, 195], sketch-to-image [106, 23], and image-to-image translation [62, 90]. Among the existing deep neural networks, the generative adversarial network (GAN) [46, 113] proved to generate reasonable, high-resolution, and realistic synthetic images [153, 90]. Indeed, its generative power ensures impressive results in many image synthesis and editing applications, e.g., image dataset augmentation [175], 3D object transformation [40], pluralistic image completion [186], or human-related image generation [118]. In particular, the latter represents one of the most interesting future research directions. In general, works performing synthesis of human-related images exploit an initial image description that usually consists of text-based or

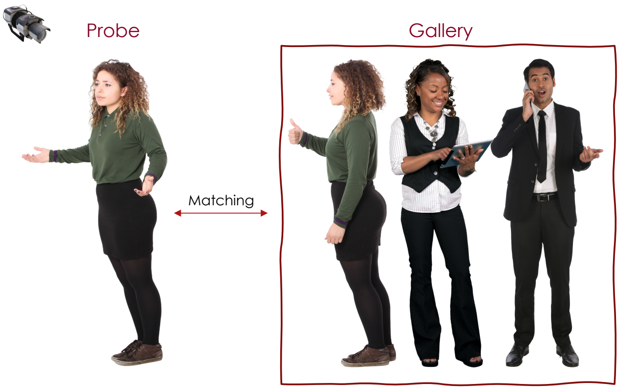


Figure 1.1. Vision-based person Re-ID example.

visual information [193, 137]. However, since also in this case traditional methods based on visual data are not helpful under challenging circumstances in which occlusions, smoke, or darkness can limit visibility, Wi-Fi signals recently have been explored to model and give a visual appearance to otherwise unavailable information carried out by radio waves [50, 68].

Wi-Fi is a mature technology that leverages radio signals transmitted by several access points (APs) to enable wireless communication between devices. When traveling between two connected devices, Wi-Fi signals are influenced by objects as well as people along their path, resulting in variations of the signal itself [25, 15]. These changes can be captured via either the received signal strength indicator (RSSI) or the channel state information (CSI) measurements [54] to implement wireless sensing applications. Despite this, the CSI is more stable and can carry more signal information with respect to the RSSI due to the underlying technology principles. In fact, for a given wireless data packet, the RSSI is represented by a single value computed at the MAC layer, indicating the relative signal quality; whereas the CSI is measured by employing the orthogonal frequency-division multiplexing (OFDM) transmission technology at the PHY layer, and includes fine-grained signal information defined at the subcarrier level [165]. What is more, this measurement has been proven to be more robust in complex environments, enabling the extraction of relevant signal characteristics, especially in indoor areas [42]. Indeed, among other things, signals amplitude and phase can be retrieved from CSI measurements and, as a matter of fact, several works exploit these radio signal properties to develop useful diverse Wi-Fi sensing applications [78, 102, 112, 155, 135, 79]. The CSI based methods are the most popular also because this measurement can be easily computed via commodity Wi-Fi devices. A sound strategy to design these sensing applications generally requires some sort of signal pre-processing to improve the received CSI measurements quality by removing, for example, amplitude outliers [27] and phase offset [139]. Subsequently, either a machine or deep learning approach is usually employed to address the given wireless sensing task [85]. Concerning the person Re-ID, systems performing person identification, i.e., methods that classify signals into known people identities, indicate that distinct people affect the Wi-Fi signals differently and, therefore, can prove particularly useful for security applications [158]. Moreover, this difference in signal variation between people was also extensively examined by previous studies on both electromagnetic absorption of human bodies and radio waves interactions with biological tissues [38, 39]; where it was shown that wireless propagation around a human body is highly dependent

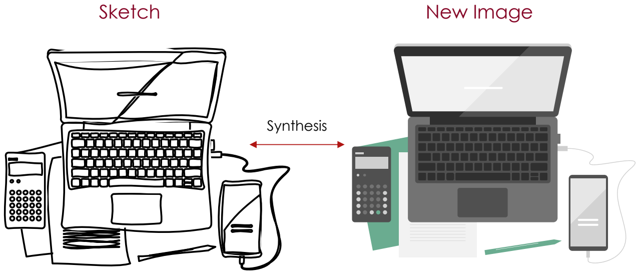


Figure 1.2. Image synthesis example from sketch.

on several characteristics such as skin and other biological tissues conditions, total body water volume, and additional physical attributes including mass and height. As a consequence, due to the high variability of such features, radio biometrics can be extracted from Wi-Fi signals to describe and ultimately recognize a given person [158]. Concerning the image synthesis, by translating radio frequency information to the visible spectrum, automatic systems can be enhanced with powerful and exciting capabilities. This domain translation is feasible because both inanimate objects and people in an environment affect the electromagnetic (EM) spectrum at different frequencies. Thus, measurements of the same entity performed at radio frequency and optic ranges can be correlated [69]. Despite this, specific mathematical operations do not exist to define this duality. However, in literature, models capable of inferring a mapping between radio and optical frequencies were introduced in the past. For instance, the ray-tracing for radio waves propagation modeling is motivated by ray optics [174]. Consequently, the analogies between the radio and visible EM spectra have made it reasonable to model the wireless signals to solve comprehension tasks, including human pose estimation synthesizing skeletal visual data [184, 185]. However, radio-based human data synthesis works are generally based on non-commodity devices or exploit raw CSI measurements. Motivated by these observations and existing literature, this thesis introduces two methods that enable person Re-ID and human-related video synthesis through Wi-Fi signals. Interestingly, the presented methods satisfy and expand the concept of a privacy-conscious system [31]; therefore, people may be re-identified and monitored by perceiving and understanding their behaviors without collecting sensitive or private information, e.g., photos or audio recordings.

For the proposed Re-ID method, starting from an estimated CSI measurement, signals amplitudes and phases are extracted and processed to improve their quality through established procedures such as the outlier removal [27], for signals amplitudes; and offset removal [139], for signals phases. Subsequently, the sanitized signals are further refined through a median filtering [59] to create a single amplitude heatmap and more stable phases along the various CSI measurements. These features, representing a person's radio biometrics, are then used as input for a novel deep neural network architecture, based on a siamese structure, that was specifically designed to extract meaningful radio biometric signatures through two parallel sub-networks inside each siamese branch. Both sub-networks are based on different models to correctly handle the computed signal amplitude heatmaps and filtered phases, i.e., a convolutional neural network (CNN) and a long-short term memory (LSTM). Afterwards, the output radio biometric signatures are used to re-identify the person across the same or two different locations. To simulate typical constrained surveillance scenarios [130] focused on single individuals, and due to the lack of existing Wi-Fi collections for the Re-ID task, a dataset was acquired capturing single persons standing between APs in several indoor locations. Moreover, comparably to real-world environments, no shielding mechanism was implemented against interference from other radio signals, since they are inevitable in our highly connected world. Thereafter, similarly to classical Re-ID methods that make use of visual information extracted from images and videos [181, 6, 192], common Re-ID metrics, such as the mean average precision (mAP) and cumulative matching characteristic (CMC), were employed to accurately evaluate the proposed methodology.

For achieving video synthesis from radio signals, instead, it is presented a novel two-branch generative neural network that maps data acquired in the wireless spectrum to the visual domain, following a teacher-student fashion [6]. Specifically, also in this case, sanitized Wi-Fi signals amplitudes extracted from CSI measurements are used to synthesize video frames depicting human silhouettes or skeletons performing different poses. The signal amplitudes were chosen since the literature proves that, other than people identity, they can discriminate human activities adequately [152, 76, 110], supporting the immediate investigation of such radio features to generate new pose-related visual data. In particular, the proposed method exploits cross-modality supervision [92] at multiple levels of the network training pipeline to transfer visual knowledge to Wi-Fi data and learn how to synthesize videos, instead of static images, of human silhouettes or skeletons from signal-based features. What is more, by leveraging 3D-GAN [148], long short-term memory (LSTM) [57], and 3D convolutional neural networks (3D-CNN)[114] architectures, the approach inherently accounts for motion information and can, therefore, also manage moving subjects. This aspect is really important to enable perception and understanding of human behavior from synthesized visual data. The latter is of particular interest since the proposed method can be helpful in typical real-world applications such as surveillance scenarios [131] by either supporting vision-based security systems or improving privacy concerns. The effectiveness of the proposed method is evaluated by performing experiments on publicly available data capturing human poses of single individuals with commodity Wi-Fi in a controlled environmental setting, typical for research-oriented and monitoring scenarios [14]. Specifically, ablation studies on the pipeline were presented via quantitative performances, based on several state-of-the-art metrics for image quality evaluation between real and synthesized video frames. In addition, a qualitative assessment through visual observations, and further investigations on the method abstraction capabilities by replacing silhouettes in the training data with skeleton frames, demonstrated the effectiveness of the proposed approach to translate radio features into a visual domain representation. Finally, as a use case example, experiments on recognizing human activities from synthesized silhouette and skeleton videos were conducted to prove the usefulness of video synthesis from wireless signals for increasing classical human activity recognition methods, such as benefiting from radio signals nature to avoid eventually vision-based complications and eliminating privacy concerns in surveillance-based applications. To this end, a small dataset was acquired by capturing single persons performing specific actions between APs in a constrained indoor environment and simulating the acquisition protocol of the aforementioned publicly available data used to evaluate the synthesis solution.

1.1 State of the Art

This section, initially, discusses the state-of-the-art regarding the use of Wi-Fi signals for solving human-related comprehension tasks. Subsequently, for both person Re-ID and image synthesis, contributions of the presented wireless sensing solutions are highlighted, also mentioning classical approaches.

1.1.1 Person Re-ID

Depending on the signal measurement type, Wi-Fi sensing methods can be broadly categorized into two classes, i.e., RSSI and CSI based approaches [73]. Concerning the RSSI measure, it indicates the received power level after any possible transmission loss, thus representing the relative signal quality. Inanimate objects (e.g., furniture) or human presence can influence radio signals and, as a matter of fact, the authors of [55] noticed significant RSSI fluctuations in both line-of-sight (LOS) and non-line-of-sight (NLOS) conditions. Supporting these findings, the RSSI signal quality was successfully employed in heterogeneous tasks such as map reconstruction [160, 66] as well as human localization [161, 37, 159, 56, 119], tracking [17] and identification [16]. Confirming the inanimate object influence on radio signals, a grid points filling with low rank matrix theory on RSSI fingerprints exchanged between several APs is used in [160], for example, to reconstruct radio maps of indoor environments; while Markov random field modeling for loopy belief propagation of sparse signals is employed, by the authors of [66], to build 3D radio maps of unknown structures using RSSI signals examined by unmanned aerial vehicles. Considering human-focused RSSI applications, instead, a popular and well-explored task is the localization one. In [161], for instance, the best AP, i.e., with the best RSSI signal quality, is selected to achieve indoor localization according to an eight-diagram approach defining the signal propagation direction; while the authors of [37] develop a feature-scaling-based k-nearest neighbors (KNN) algorithm and further refine the RSSI signals via outlier removal to address an analogous task. Further improvements to localization systems are also provided by techniques that can clean up the received signals. For example, in [159, 56] and [119], Gaussian, weighted average, and continuous wavelet transform (CWT) filtering are applied, respectively, to improve the input for the chosen localization algorithms. The received signal quality increment can also be obtained by reducing possible interferences as shown in [17], where a custom communication protocol enables to track humans via RSSI measurements; or by using wearable devices as demonstrated by the authors of [16], that present an RSSI proximity algorithm able to identify several persons. What is more, improved RSSI measurements are also successfully used to address other complex tasks such as human action recognition. Indeed, as described in [48], sanitizing the RSSI through outlier removal and Gaussian filtering, enables a feature fusion approach to obtain significant results on the action recognition task, therefore indicating that the Wi-Fi technology is a good medium for sensing applications and could be also effectively applied to other tasks.

Regarding the CSI measure, it captures richer information about the wireless transmission among communicating APs, contrary to the RSSI that does not provide fine-grained features except from the relative signal quality of a wireless environment. For instance, CSI can acquire amplitude and phase features for each subcarrier in the OFDM channel [194]. By describing the channel characteristics of a frequency-diverse group of subcarriers, this measurement is more robust to narrowband interference from other signals. Moreover, it is not affected by the automatic power level adjustment algorithm implemented in commodity wireless APs [29]. As a consequence, the channel state information is gaining momentum in the latest years. In [141], for example, the authors show that there is a high correlation between subsequent CSI measurements, and consistently detect falling humans via CSI amplitude. Similarly, [93] and [135] detect falling humans through denoised frequency spectrogram images, in the former, and phase differences, in the latter, that can be both extracted from CSI measurements; thus indicating that CSI is an information-rich measure. Indeed, among others, Wi-Fi signals amplitude [27, 117], phase [139, 127], and frequency [156], obtained from CSI, are also effectively employed in other tasks such as human indoor localization. For example, the authors of [117] apply a fingerprint matching procedure after optimizing a centroid-based algorithm (i.e., KNN) used to examine locations through CSI amplitudes. In [139], instead, indoor locations fingerprints are defined by linearly transformed CSI phases, where offsets deriving from the transmission are removed; while in [156] the indoor localization is addressed through passive radio maps that are analyzed via a probabilistic algorithm detecting anomalies in CSI frequencies. Furthermore, CSI measurements can also be exploited to capture human movements and, consequently, perform action recognition from the received signals. For instance, the authors of [41] use several CSI channels to produce radio images that enable both to localize and to recognize the corresponding human actions. Similarly, in [151], activity recognition is achieved through variance-normalized CSI amplitude waveforms filtered by the principal component analysis (PCA) procedure. In [163], instead, gestures are recognized via a spatiotemporal examination of CSI phases executed by a siamese recurrent neural network; an architecture that extracts meaningful features from the input phases, and which has been extended in the proposed approach since it is already successfully applied in classical vision-based re-id approaches [150, 187, 26, 95, 126].

Although CSI-based works generally tend to focus on a single Wi-Fi signal characteristic (e.g., amplitude), it is not uncommon to find approaches exploiting more information derived from CSI measurements. Such an example is found in [49], where both CSI amplitude and phase, computed along two distinct RX devices, are used jointly to generate fine-grained human skeleton poses. Indeed, the multiple CSI channels can provide relevant information about people between two, or more, APs and, as a matter of fact, they are employed to define structural biometric features in [158]. In particular, these features represent body pose differences that can be registered and distinguished through the CSI by employing a time-reversal (TR) technique, that ultimately enables to identify a known person. Following a similar reasoning, the authors of [74] perform user identification through CSI frequency shifts associated to gestures, by band-pass filtering the signals and further refining them through PCA. In [30], instead, CSI amplitudes, treated with discrete wavelet transforms (DWT) and statistical profiling (e.g., channel power distribution), are coupled with people's gait allowing for their identification through Wi-Fi signals. In addition, further confirming the CSI measurement effectiveness, filtered CSI amplitudes are exploited in [63] to simultaneously learn several tasks such as action recognition, user tracking and identification, through a deep neural network. Finally, in [133], the authors are able to also extract internal characteristics (i.e., respiration rates) from CSI measurements by improving the SNR of signals associated to breaths, which allow to identify known users; indicating that the CSI measure contains different signal cues for distinct people, and therefore supporting our direct investigation of CSI measurements to describe a person through radio biometric signatures to address the re-identification task.

1.1.2 Image Synthesis

The generative adversarial learning has been widely used for image synthesis in several fields during the last few years. In [62], the authors address the paired image-to-image translation investigating conditional GAN networks. They observed that conditioned generative modeling efficiently handles tasks requiring photographic output. Indeed, also in [178] two stacked conditional GAN models are used for photo-realistic image synthesis from text data. However, they divide the text-to-image problem into two main stages. The Stage-I GAN generates low-resolution images while the Stage-II GAN, stacked on top of it, generates realistic high-resolution images conditioned by the previous stage results and text descriptions. Once again, GANs are also used in [103] for cross-view image synthesis to produce, principally, outdoor scene images combining aerial and street image views. The authors of [106] focus on the GAN-based sketch-to-image synthesis method, constraining the generative process with sketched boundaries and color strokes to obtain realistic images. Instead, in [196], the authors use cycle-consistent adversarial networks (CycleGANs) for image-to-image translation but, differently from [62], learn to translate an image from a source domain X to a target domain Y in the absence of paired training examples. Recently, the GAN is being increasingly used to synthesize human-related images. For instance, in [167] remarkable results are achieved for person search in video sequences by combining the GAN capability with a deep complementary classifier based on a convolutional neural network (CNN). They used the GAN to generate new samples for the training set augmentation, improving the classifier performance during the testing phase. In [32], instead, authors present an identity-aware CycleGAN for face photo-sketch synthesis and recognition. They improve the CycleGAN performance to solve the photo-sketch synthesis task by giving attention to the facial regions, including eyes and nose, which can be significant in identity recognition. Focusing on the dynamic synthesis of facial expressions, in [45] is introduced a GAN working with a series of semantic parts with different shapes to describe geometrical facial movements. For person monitoring applications, instead, the GAN is primarily used to improve deep models performances [189, 83, 176]. On a different note, a model trained on a specific dataset does not work when applied on another collection. The authors of [190] solve such an issue by proposing a GAN-based image-to-image translation, transferring the images of a source dataset to the style of each camera in a target set of data. Similarly, in [84] is proposed a GAN structure to preserve features for cross-domain person re-identification, solving problems arising from significant variations between the training dataset and the target scene.

The state-of-the-art validates the incredible power of generative adversarial learning for image synthesis starting from text, sketch, or another image, as the information source. However, nowadays, the Wi-Fi signal is being explored for synthesizing visual data, opening up a new frontier for image synthesis and surveillance applications. Indeed, in [134], the authors propose a method for person perception using CSI measurements. They combine multiple residual convolution blocks and U-Net [105] models to synthesize either human silhouette or pose. Similarly, only for the latter, in [69] is introduced a network architecture comprising three CNN-based sub-models: CSI encoder, domain translator, and frame decoder. The first encodes the CSI measurements, building the latent representations of radio features; the second connects such representations to the visual domain. Finally, the third generates images from the information in the latent visual domain. Again, the authors of [50] propose a two-branch network to synthesize the human skeleton from a CSI-based image. The top branch leverages the OpenPose [19] framework to extract, from video frames, the skeletal data used as supervision. Instead, the bottom branch generates the corresponding skeleton for each CSI image exploiting the supervision data as ground truth (GT). A similar supervision strategy is followed in [68] for image synthesis of stationary subjects starting from the signal-to-noise ratio (SNR) of the Wi-Fi signal. They perform an ablation study on different classical GAN-based structures combined with a complementary detection system to quantitatively evaluate their synthesis capabilities. Despite the remarkable results, none of the tested methods considers the motion aspect of moving subjects. To address the latter, the proposed method improves the generative adversarial learning strategy by exploiting the 3D-GAN architecture. Moreover, the model generalization capabilities are further improved by manipulating the middle-level representations used to transfer visual knowledge to wireless signal-based features, ultimately enabling for the human silhouette and skeleton video synthesis via radio-to-visual domain translation.

1.2 Contribution

This section reports in brief the contributions of this thesis with respect to the state-of-the-art. Concerning the person Re-ID, the primary feature is the definition of a completely new Wi-Fi sensing application by implementing the Re-ID based exclusively on Wi-Fi signals that can also avoid classical vision-based drawbacks thanks to the different medium nature of the wireless technology. This is done by designing a novel architecture based on a siamese model structure, leveraging parallel sub-networks in each siamese branch to extract meaningful radio biometric signatures from radio signals examined at either the same or at different locations. Concerning the image synthesis, the design of a novel two-branch generative Wi-Fi sensing framework that inherently considers motion information to synthesize coherent human silhouette and skeleton videos from wireless signals is proposed. Moreover, through CSI measurements noise removal, a Wi-Fi data sanitization procedure is applied to obtain robust radio features based on signals amplitudes and increase the synthesized video quality.

Finally, to test the person Re-ID approach, a specific dataset was acquired to prove the effectiveness of signal-based methods to address the Re-ID task in typical constrained surveillance scenarios. Instead, publicly available Wi-Fi data focused on humans performing different continuous poses were used to evaluate the image synthesis solution in controlled environmental settings typical for monitoring applications. In addition, a small dataset on human activities was acquired to show the benefit of using video synthesis from wireless signals for enhancing traditional human activities recognition systems capabilities, avoiding well-known vision-related issues and eliminating privacy concerns in people monitoring-based applications.

Chapter 2

The wireless channel

This chapter introduces the theoretical insights about the wireless communication channel required to understand the underlying technology of the Wi-Fi sensing applications, including person Re-ID and video synthesis through radio signals.

2.1 Physical Model for Wireless Channel

The main characteristic of all wireless communications, including Wi-Fi transmission, is the use of electromagnetic (EM) radiation to convey information from a transmitting (TX) to a receiving (RX) antenna. Such radiation travels at a constant speed leveraging a sinusoidal waveform propagated through an unguided transmission medium, e.g., air or space, as shown in Fig. 2.1a. Each EM wave, which results from oscillations between electric and magnetic fields producing an electromagnetic field, is characterized by wavelength and frequency properties. The former, shown in Fig. 2.1b, is the waveform length between two adjacent crests or troughs of the wave and is measured in meters (m). The latter, shown in Fig. 2.1c, is the number of oscillation cycles per second and is measured in Hertz (Hz). There exists a relation between the wave propagation velocity v, wavelength λ, and frequency f, that can be expressed as follows:

$$v = \lambda f, \tag{2.1}$$

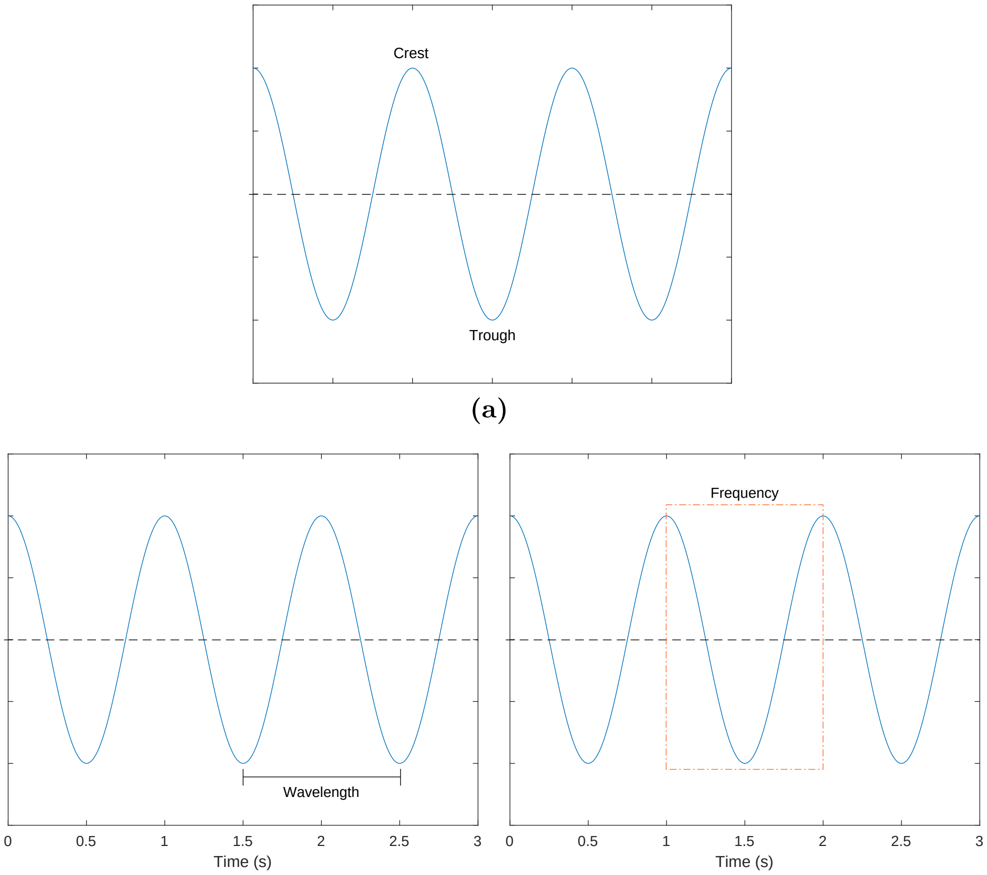
where v is the speed of light c in a vacuum, i.e., 3×108 m/s , or less in a different transmission medium. Therefore, starting from Eq. (2.1), λ and f can be defined as:

$$\lambda = \frac{v}{f}, \tag{2.2}$$

$$f = \frac{v}{\lambda}, \tag{2.3}$$

showing that the frequency and wavelength of an EM wave are directly proportional to the propagation velocity and inversely proportional to each other. The electromagnetic field generated by traveling radiation changes depending on the distance from the transmitting source, as shown in Fig. 2.2. Practically, the field defined as one wavelength or less from the TX antenna is called near-field. Within this region, the strength of radiation quickly decreases as the wave moves away from its source and a close receiver influences the transmitter radiation. Indeed, this electromagnetic field region is employed to implement NFC communication, where it is required that the receiver NFC tag is located within 4 cm from the transmitter. However, a different electromagnetic field region starts two wavelengths from the TX antenna and propagates outward; this is called the far-field. Differently from near-field within this region, traveling away from the transmitting source, the power of the electromagnetic field decreases inversely to the square of the distance from the TX antenna, and a receiver does not influence the original wave radiated by a transmitting antenna. Due to its properties, the far-field enables the long-range transmissions required for most wireless sensing applications that employ the widely used radio communication methods, e.g., Bluetooth or Wi-Fi.

For the definition of a physical model for a wireless channel [128], the starting point is a fixed TX antenna emitting a wave into free space. Being both electric and magnetic fields proportional to each other, other than perpendicular to one another and the direction of the EM wave propagation, it is sufficient to estimate only one of them to describe the far-field of the transmitting source. Given a sinusoid $cos\ 2\pi ft$ transmitted in a vacuum, in the absence of an RX antenna, the responding electric



(a)

(b)

(c)

Figure 2.1. Electromagnetic radiation properties. In (a) the waveform representation, in (b) the wavelength, and in (c) the frequency per second delimited by the orange dashed rectangle.

far-field at time t can be modeled as follows:

$$E(f, t, (r, \theta, \psi)) = \frac{\alpha_s(\theta, \psi, f)\ cos\ 2\pi f(t - r/c)}{r}, \tag{2.4}$$

where (r, θ, ψ) is the point in the space on which the electric field is measured, r is the distance from the transmitter, θ and ψ are the horizontal and vertical angles from the TX antenna to the considered point in the space, respectively. Due to the vacuum, c indicates the speed of light as propagation velocity, and $\alpha_s(\theta, \psi, f)$ is the far-field radiation pattern of the transmitting device in the direction (θ, ψ) at the specific frequency f. The term α_s is a scaling factor to take into account antenna losses. The radiation pattern of an antenna indicates the variation of the spatial distribution of power of the radiated EM wave as a function of angular direction, and a graphical example is shown in Fig. 2.3. Each pattern is divided into radiation lobes, and the main lobe is in the direction of maximum radiation power. Instead, the side lobes are undesired radiation directions characterized by less power. Finally, the back lobe is the direction with minimum radiation power. As a consequence of the reciprocity theorem of electromagnetics, the sensitivity of an antenna as a function of angular direction when used as the receiver, i.e., the receiving pattern, is equal to the far-field radiation pattern of the same antenna when used as a transmitter [122].

Following, considering the presence of a fixed RX antenna at a specific location $\mathbf{x} = (r, \theta, \psi)$, without noise, the waveform received in response to the same sinusoid can be linearly defined as:

$$E_r(f, t, \mathbf{x}) = \frac{\alpha(\theta, \psi, f)\ cos\ 2\pi f(t - r/c)}{r}, \tag{2.5}$$

where $\alpha(\theta, \psi, f)$ is the product of transmitting and receiving patterns of TX and RX antennas in the given direction (θ, ψ). Adding the RX antenna, the electric field changes in the proximity of $\mathbf{x}$ and this is considered by its receiving pattern. Because of both Eqs. (2.4) and (2.5) are linear in the input; the received EM field as a response to a weighted sum of transmitted waveforms at $\mathbf{x}$ is the weighted sum of responses to each one of those waveforms. Then, for a given location $\mathbf{x}$, we can define:

$$H(f) := \frac{\alpha(\theta, \psi, f)\ e^{-j2\pi fr/c}}{r}, \tag{2.6}$$

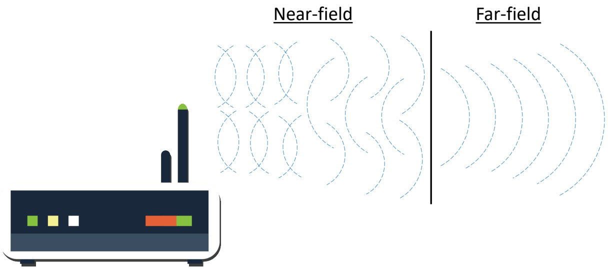


Figure 2.2. The two different electromagnetic field regions: near-field and far-field.

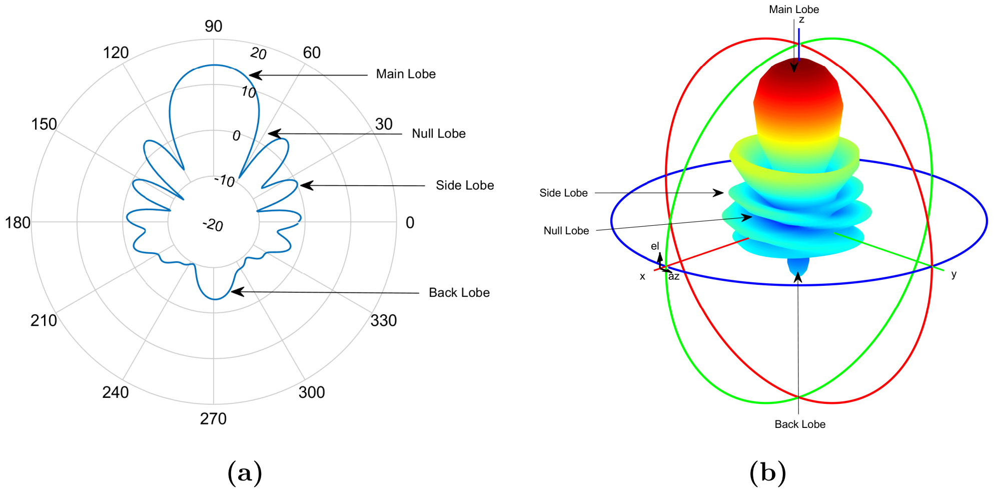


Figure 2.3. Example of antenna radiation pattern in 2D (a) and 3D (b) plots for a comprehensive view.

having that $E_r(f, t, \mathbf{x}) = \Re[H(f)e^{j2\pi ft}]$. Therefore, $H(f)$ is the definition of Eq. (2.5) as the system function of a linear time-invariant channel in frequency domain and its inverse Fourier transform is the corresponding impulse response in time domain. However, the time-invariant property cannot be considered if the antennas or possible obstructions are in relative motion. Therefore, considering the fixed TX antenna and free space model defined by Eq. (2.4), examining a specific point moving at velocity v in the direction of increasing the distance r from the wave transmitter, expressed as $r(t) = r_0 + vt$, the responding electric far-field at time t is defined as:

$$E(f, t, (r_0 + vt, \theta, \psi)) = \frac{\alpha_s(\theta, \psi, f)\ cos\ 2\pi f(t - r_0/c - vt/c)}{r_0 + vt}, \tag{2.7}$$

where $f(t - r_0/c - vt/c)$ can be rewritten as $f(1 - v/c)t - fr_0/c$ focusing on the Doppler shift [43] of $-fv/c$ caused by the motion of the observed point moving away from transmitting source. Therefore, the EM wave at frequency f changes to a radiation of frequency $f(1 - v/c)$. This means that every consecutive crest in the transmitted wave moves a little further before being captured at the examined point. Subsequently, considering the model defined by Eq. (2.5), if the RX antenna is placed at a moving location defined as $\mathbf{x}(t) = (r(t), \theta, \psi)$, again with $r(t) = r_0 + vt$, the received waveform is determined as follows:

$$E_r(f, t, \mathbf{x}(t)) = \frac{\alpha(\theta, \psi, f)\ cos\ 2\pi f[(1 - v/c)t - r_0/c]}{r_0 + vt}. \tag{2.8}$$

The Eq. (2.8) cannot be described as a system function of a linear time-invariant channel as it expressed. However, ignoring the time-varying distance $r(t)$ at the denominator, we can still define the channel as a system function by considering the Doppler shift $-fv/c$ rather than the corresponding frequency f.

2.2 Linear System for Wireless Channel

In the previous section, we defined the wireless channel as the response to a sinusoidal input $\phi(t) = cos\ 2\pi ft$ that is propagating into free space observing a fixed and moving receiving antenna. However, in real scenarios, it is necessary to account for radio waves propagation delay caused by the total distance traveled at the transmission medium speed. In addition, during the signal radiation, should also be considered different attenuation factors causing the wireless channel strength variations over both time and frequency due to large- and small-scale fading [128]. The former, typically frequency independent, is caused by the path loss of the signal as a function of the distance and the shadowing effect by significant obstacles along the propagation path, e.g., buildings. The latter, typically frequency-dependent, is caused by the constructive and destructive interference due to the multi-path effect that makes several signal replicas reach the receiver across multiple paths interfering in amplitude and phase. Therefore, following these observations, a multi-path wireless channel at specific frequency f can be linearly expressed as:

$$y(f,t) = \sum_i a_i(f,t)\phi(t - \tau_i(f,t)), \tag{2.9}$$

where $y(f,t)$ is the received signal as response to a transmitted sinusoid $\phi(t)$, $a_i(f,t)$ and $\tau_i(f,t)$ are the signal attenuation and propagation delay, respectively, at time t on the i-th path between TX and RX antennas. Assuming that individual path attenuation and propagation delay are independent from frequency, for the principle of superposition, Eq. (2.9) can be simplified as follows:

$$y(t) = \sum_i a_i(t)x(t - \tau_i(t)), \tag{2.10}$$

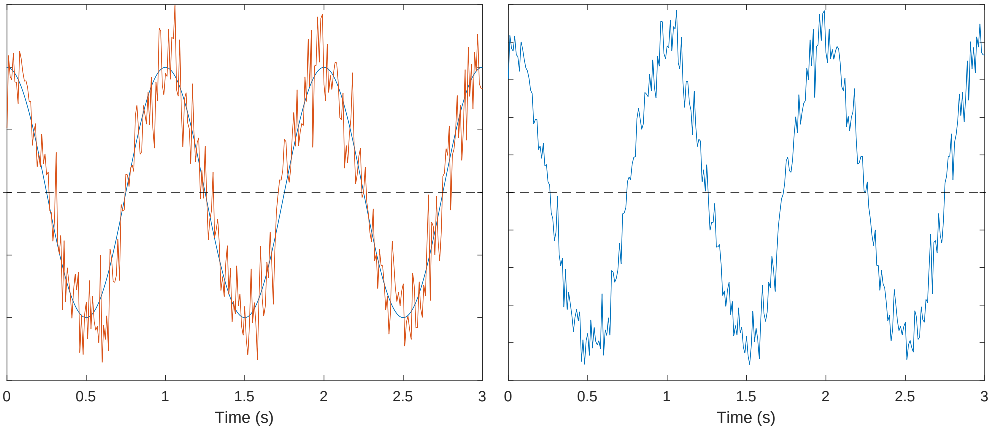
with $x(t)$ an arbitrary transmitted bandpass signal. Despite the previous frequency independence assumption on individual paths, notice that the channel response can change according to frequency because different paths with different delays in multi-path propagation cause multi-path fading. Being the Eq. (2.10) linear, a fading multi-path wireless channel can be modeled, in time domain, as the channel impulse response (CIR) at time t of a linear time-variant channel, as follows:

$$h(\tau,t) = \sum_i a_i(t)\delta(\tau - \tau_i(t)), \tag{2.11}$$

where $a_i(t)$ and $\tau_i(t)$ still be the attenuation factor and propagation delay of the i-th path, respectively, while $\delta(\tau)$ corresponds to the Dirac delta function. Therefore, comparing the Eqs. (2.9) and (2.11), the relation between the received signal, at time t, and the bandpass signal x, transmitted at time $t - \tau$, can be defined in terms of impulse response as:

$$y(t) = \int_{-\infty}^{\infty} h(\tau,t)x(t - \tau)d\tau. \tag{2.12}$$

Applying the Fourier transform (FT) to the impulse response, it is possible to estimate the corresponding time-varying channel frequency response (CFR) defined



(a) **(b)**

Figure 2.4. Example of noisy signal. In (a) noise-free component in blue, and AWGN component in orange. In (b) the resulting signal affected by AWGN noise.

as:

$$H(f;t) = \int_{-\infty}^{\infty} h(\tau,t)e^{-j2\pi f\tau}\, d\tau = \sum_i a_i(t)e^{-j2\pi f\tau_i(t)} = |H(f;t)|e^{j\angle H(f;t)}, \quad (2.13)$$

where $|H(f;t)|)$ and $\angle H(f;t)$ indicate the signal amplitude and phase responses, respectively, and j is the imaginary component. Finally, in the frequency domain, a time-variant channel can be linearly modeled as:

$$y(t) = H(f;t)\; x(t-\tau). \quad (2.14)$$

However, notice that in the case of static environment with all fixed TX and RX antennas, the wireless channel can be described as a linear time-invariant system with Eqs. (2.11) and (2.13) specified as follows:

$$h(\tau) = \sum_i a_i\; \delta(\tau - \tau_i), \quad (2.15)$$

$$H(f) = \int_{-\infty}^{\infty} h(\tau)e^{-j2\pi f\tau}\, d\tau = \sum_i a_i\; e^{-j2\pi f\tau_i} = |H(f)|e^{j\angle H(f)}, \quad (2.16)$$

where attenuations a_i and propagation delays τ_i do not change with time t. As the last step, in the definition of the wireless communication channel as a linear system, other than factors causing signal attenuation and propagation delay, there is the need to consider factors modeled as random and referred to as noise. It is common for all communication systems in the real world to face such a noise effect, shown in Fig. 2.4; indeed, it is physically impossible to have a noise-free channel. From the signal formal definition point of view, the additive white Gaussian noise (AWGN) is usually chosen for noise modeling on the received signal. Following this observation, the wireless channel expressed by Eq. (2.10) under noisy conditions can be defined as:

$$y(t) = \sum_i a_i(t)x(t-\tau_i(t)) + \omega(t), \quad (2.17)$$

where $\omega(t)$ is the AWGN component at time t. The noise component is additive because, in the model specified by Eq. (2.17), the RX antenna process the sum between a noise-free component, i.e., the signal that would have been received in the absence of noise, and the noisy component independent from the transmitted signal. Such component is random and, at each time, drawn from a fixed zero-mean Gaussian distribution. The Gaussian is chosen because the noise results from adding the effects of several and independent factors allowing the application of the central limit theorem [35]. This theorem establishes that the sum of independent random variables tends toward a normal distribution. Concluding, the noise is named white because it has a uniform power across the entire frequency band.

2.3 Received signal strength indicator

The signal received from an RX antenna is characterized by a reduction in the intensity of the waveform power accumulated as the EM radiation propagates away from the transmitter and encounters several obstacles in the environment. The received signal strength indicator (RSSI) indicates the relative signal quality capturing the wireless channel strength variations. Indeed, such a measurement has proved suitable for assessing the link quality in wireless networks [132] and developing diverse Wi-Fi sensing applications [17, 56, 64]. Formally, the RSSI is measured in decibels (dB) and generally expressed following the log-distance path loss model [3], defined as:

$$PL_d = PL_{d_0} + 10n \log_{10} \frac{d}{d_0} + \chi, \tag{2.18}$$

where PL_d is the path loss at an arbitrary distance d, n is the path loss exponent depending from the environment, χ is a zero-mean Gaussian distributed random variable with standard deviation σ in dB scale, representing noise to model shadowing variations. Finally, PL_{d_0} is the path loss at a reference distance d_0 usually computed

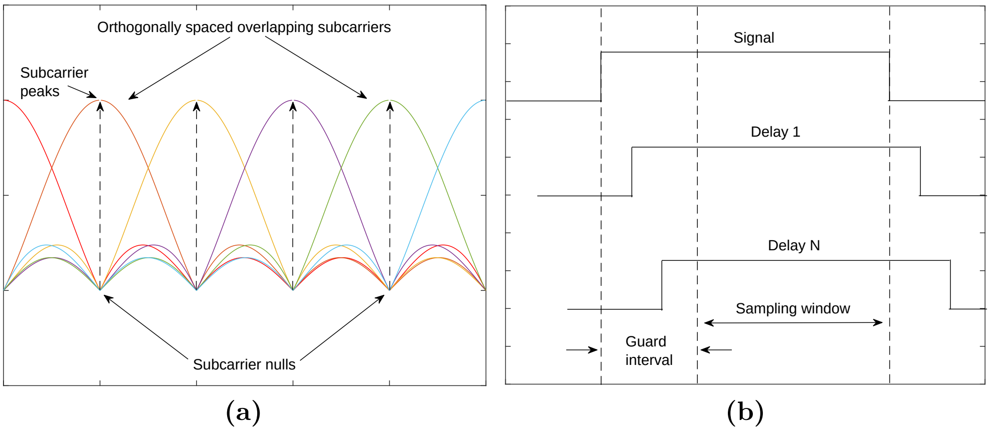


Figure 2.5. OFDM technology simplified graphical representations. In (a) OFM signal modulation example, in (b) guard interval visual interpretation.

by exploiting the free-space path loss model, derived from Friis transmission equation [36], defined as:

$$FSPL = 10\log_{10}\left(\frac{4\pi df}{c}\right)^{2}, \tag{2.19}$$

where d is the distance between TX and RX antenna, c is the speed of light, and f is the radio wave frequency obtained dividing the speed of light by the wavelength.

2.4 Orthogonal Frequency-Division Multiplexing

Most modern wireless communications, including IEEE 802.11a/b/g/n/ac/ax Wi-Fi systems, employ the orthogonal frequency-division multiplexing (OFDM) signal modulation [21, 144] to transmit data encoded on multiple orthogonal carrier frequencies, i.e., subcarriers, within the same individual communication channel. Specifically, as shown in Fig. 2.5(a), a high-rate data stream is transmitted using closely spaced orthogonal subcarriers carrying information in parallel, each of which modulated using a digital low symbol rate existing modulation scheme. By applying any modulation scheme to a carrier, sidebands result from such a process causing the overlap between different subcarriers involved in a multiple carriers scenario, creating interference at any overlap frequency. However, this is not valid for orthogonal frequencies where the orthogonality property generates signal nulls among adjacent subcarriers to which each subcarrier peak is aligned. Therefore, the receiver can recover the original signal correlating the known set of orthogonal sinusoids without interference despite the overlapping sidebands.

Furthermore, combining various carrier frequencies allows reaching total data rates comparable to single-carrier based modulation techniques within the same bandwidths while preventing interference or signal corruption eventually caused by the multi-path effect. The latter is possible using the guard intervals, ensuring that new delayed replicas of the signal received do not alter the timing and phase of the signal itself because data is only sampled in a stable condition as shown in Fig. 2.5(b). Therefore, the OFDM technique improves the classical frequency division-multiplexing approach, exploiting multiple carriers to convey the information within a channel, orthogonality to prevent interference caused by overlapping frequencies, and guard bands for signal stability.

2.5 Channel State Information

The channel state information (CSI) is an alternative and more fine-grained wireless communication channel measurement than the RSSI metric described in Sec. 2.3. It is widely used in modern wireless communications based on the OFDM system to obtain detailed signal propagation characteristics from the transmitter to the receiver at the subcarrier level. In this way, transmissions can be adapted to the current link conditions enabling reliable communications in multiple-input multiple-output (MIMO) antenna configuration [109] shown in Fig. 2.6. While the RSSI indicates the relative quality of the received signal, the CSI can capture richer information about the signal such as amplitude, phase, or frequency. Indeed, as a consequence, such a measurement is gaining momentum in recent years for developing numerous

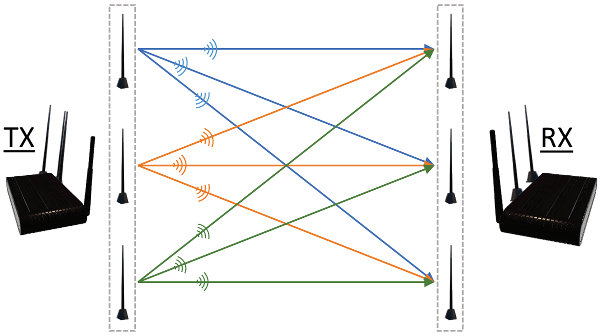


Figure 2.6. Example of MIMO antenna configuration. Multiple TX antennas transmit multiple parallel signals at the same frequencies to multiple RX antennas.

Wi-Fi sensing applications, including human presence detection [44, 146], activity recognition [138, 154], and tracking [97, 177]. In a standard Wi-Fi transmission, P data packets characterize the signal exchanged between TX and RX access points (APs). The CSI is a frequency-domain evaluation measure involving the CFR values computed for all the K OFDM-based subcarriers related to each $p \in P$ packet reaching the receiver. Given Θ and Γ arrays of fixed receiving and transmitting antennas placed in a static environment, respectively, for each subcarrier $k \in K$ over the wireless communication established between the $\theta \in \Theta$ and $\gamma \in \Gamma$ antennas, the frequency response $H(f)_k^{\theta,\gamma}$ can be specified as:

$$H(f)_k^{\theta,\gamma} = |H(f)_k^{\theta,\gamma}| e^{j \angle H(f)_k^{\theta,\gamma}}, \tag{2.20}$$

where $|H(f)_k^{\theta,\gamma}|$ is the signal amplitude, $\angle |H(f)_k^{\theta,\gamma}|$ is the signal phase, and j the imaginary component resulting from the Fourier transform applied on the impulse response. Therefore, the final CSI measurement estimated over all the K subcarriers, considering all the antennas in the TX and RX arrays, is a complex matrix of size $\Theta \times \Gamma \times K$, defined as:

$$CSI = \begin{bmatrix} H(f)_1^{(1,1)} & H(f)_2^{(1,1)} & \dots & H(f)_\kappa^{(1,1)} \\ H(f)_1^{(1,2)} & H(f)_2^{(1,2)} & \dots & H(f)_\kappa^{(1,2)} \\ \vdots & \vdots & \vdots & \vdots \\ H(f)_1^{(\theta,\gamma)} & H(f)_2^{(\theta,\gamma)} & \dots & H(f)_\kappa^{(\theta,\gamma)} \end{bmatrix}, \tag{2.21}$$

where $H(f)_\kappa^{(\theta,\gamma)}$ is a signed 8-bit complex number indicating the κ-th subcarrier CFR value over the $\theta \in \Theta$ and $\gamma \in \Gamma$ antennas.

2.6 Signal Amplitude

One of the most common signal characteristics derived from the CSI measurement specification is its amplitude. During the propagation, the radio wave generates patterns of disturbance oscillating through the transmission medium; therefore, theoretically, such a measurement is the maximum displacement of points on the

EM wave measured from its equilibrium position, as shown in Fig. 2.7. Practically, instead, it can express the power of the radio signal reported in the dB scale. In real scenarios, the wireless channel is not a noise-free communication system, and it is accordingly modeled as in Eq. (2.17). Therefore, notice that even if the amplitude patterns can be retrieved from the CSI matrix, they require further processing to alleviate the possible noise from wireless protocol specifications and ambient conditions. The solution can be the use of outliers removal strategies [98, 27].

2.6.1 Amplitude Sanitization

Abnormal values can occur in the CSI measurement and, consequently, influence the signal amplitude extraction procedure; therefore, such outliers should be removed. Sanitizing the amplitude through a filtering-based strategy allows eliminating eventually irrelevant radio information not necessarily required to solve the addressed Wi-Fi sensing task, namely, to mitigate noise caused by various factors such as furniture material and position or other external radio interference, as shown in Fig. 2.8(a). To this end, it is possible to investigate the Hampel identifier [28] to identify an outlier value as any point falling outside a closed interval $[\mu - \xi\sigma, \mu + \xi\sigma]$, where μ and σ are the median and median absolute deviation (MAD) of the data sequence, respectively, and ξ is an application-dependent constant. Precisely, given a sliding window of fixed length, outliers are identified by points resulting in more than ξ local MAD away from the local median within this window over the wireless data packets for each subcarrier. First, the local outliers are detected by exploiting local median values over the sliding window. Afterwards, the detected abnormal values are replaced with the previous non-outlier number to maintain consistent amplitude information.

Formally, given the signal amplitudes extracted from the CSI measurements of $p \in P$ wireless packets transmitted between the TX and RX antennas, and considering the size of a window w, the local median is defined as follows:

$$\mu(\Omega^{p,\kappa}) = \Omega^{p,\kappa}_{\lceil w/2 \rceil}, \tag{2.22}$$

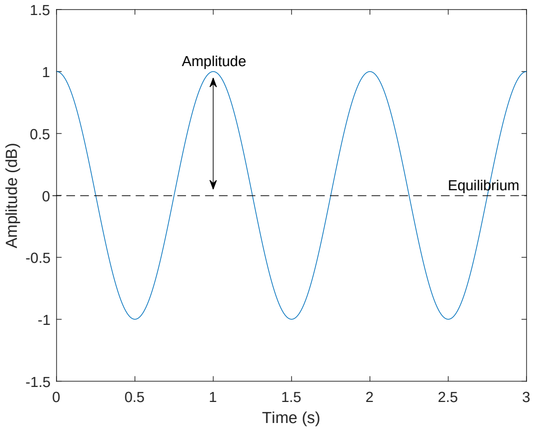


Figure 2.7. Example of EM wave amplitude.

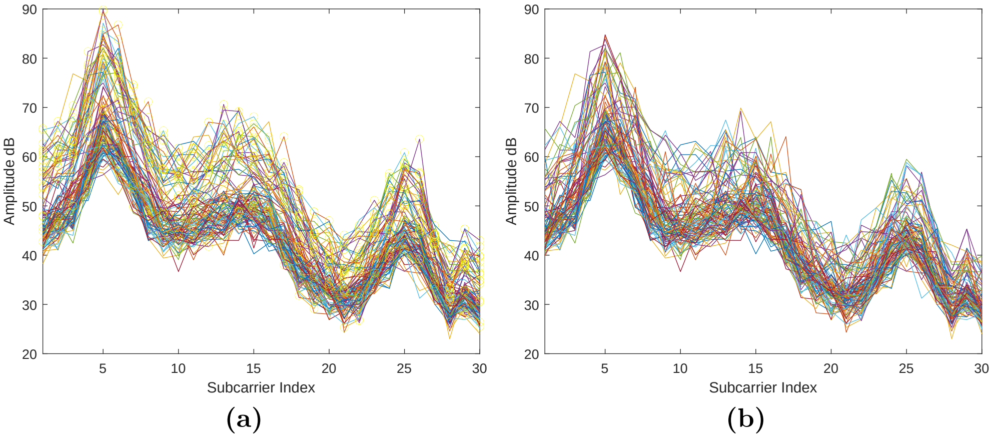


Figure 2.8. CSI extracted amplitudes processing example. In (a) and (b) the raw and sanitized amplitudes for one TX-RX antenna pair, respectively. Yellow circles are abnormal values in raw data.

$$\begin{aligned}\Omega^{p,\kappa} = \Big\{&|H(f)_\kappa|^{p-\lfloor w/2\rfloor}, \dots, |H(f)_\kappa|^{p+\lfloor w/2\rfloor} : \\ &|H(f)_\kappa|^{p-\lfloor w/2\rfloor} < |H(f)_\kappa|^{p+\lfloor w/2\rfloor}\Big\},\end{aligned} \tag{2.23}$$

where $\Omega^{p,\kappa}$ is the set containing w neighboring packets amplitude $|H(f)_\kappa|$ of the κ-th subcarrier, in ascending order. Notice that we describe the equation for a single sample for the sake of simplicity; however, μ is computed over all $\Theta \times \Gamma \times K$ antennas and subcarriers combinations. Therefore, the local MAD used to detect abnormal amplitude values is defined as:

$$\begin{aligned}\sigma(\Omega^{p,\kappa}) = \; &\mu(|\Omega_i^{p,\kappa} - \mu(\Omega^{p,\kappa})|), \\ &\forall i, \text{ s.t. } 1 \leq i \leq w.\end{aligned} \tag{2.24}$$

Finally, the intervals in which points are acceptable local values are defined as:

$$limit^{p,\kappa} = \mu(\Omega^{p,\kappa}) \pm \xi * \sigma(\Omega^{p,\kappa}), \tag{2.25}$$

and each value falling outside these ranges is replaced with the previous non-outlier value to maintain information consistency, obtaining sanitized amplitudes as shown in Fig. 2.8(b).

2.7 Signal Phase

Another common signal characteristic computed starting from the CSI measurement is the signal phase. Such a measure is related to two or more radio signals sharing the same frequency at a reference time. Specifically, given a point in time t, multiple waves with the same frequency are said to be in phase if they are completely aligned, whereas if crests and troughs do not overlap precisely, the radio waves are said to be out of phase. Therefore, the phase can express the relative displacement among the same frequency radio signals expressed in degrees or radians. Fig. 2.9 depicts an original signal and its shifted version in time, causing the phase offset β. It is

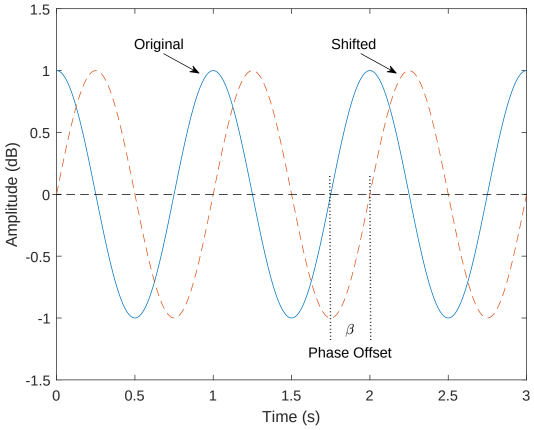


Figure 2.9. Example of a signal out of phase. The blue sinusoidal wave is the signal originally transmitted, the dashed orange sinusoid is a delayed version of the same signal, i.e., shifted in time, causing phase offsets.

crucial to notice that the phase influences the received radio wave amplitude. If the signals are in phase, the combined amplitudes increase the received signal strength; instead, if the signals are 180° out of phase, the resulting signal strength is null because crests of one wave are aligned with troughs of the other wave canceling each other. Similar to the amplitudes, phases can be retrieved from the CSI matrix but require further processing to develop Wi-Fi sensing applications [99, 139].

2.7.1 Phase Sanitization

The signal phase behaving utterly random due to random noise, unsynchronized time clocks between transmitting and receiving antennas, or the Doppler effect caused by the communicating devices relative motion, makes the raw phase information completely unusable and leads to random phase offsets, as shown Fig. 2.10(a). The CSI extracted phases can be calibrated by applying a linear transformation based technique as recommended in [108, 139] to address and mitigate such a problem.

Formally, the raw CSI phases $\angle\widehat{H}(f)_k$ measured for the k-th subcarrier can be expressed as:

$$\angle\widehat{H}(f)_\kappa = \angle H(f)_\kappa + 2\pi\frac{m_\kappa}{N}\Delta t + \beta + Z, \tag{2.26}$$

where $\angle H(f)_\kappa$ is the real phase, Δt is the timing offset at the receiver corresponding to the time interval between the signal arrival and detection, β is the unknown phase offset, Z is the measurement noise, while m_k and N correspond to the subcarrier index and fast Fourier transform (FFT) size as specified by the IEEE 802.11n standard [53]. Since Δt and β are unknown, the genuine phase information cannot be directly retrieved. However, considering the phase across the entire frequency band, the unknown terms can be removed through a linear transformation, mitigating the effect of random noise. To this end, being the subcarrier frequencies symmetric, the phase slope a and offset b over the total frequency band can be represented via the following equations:

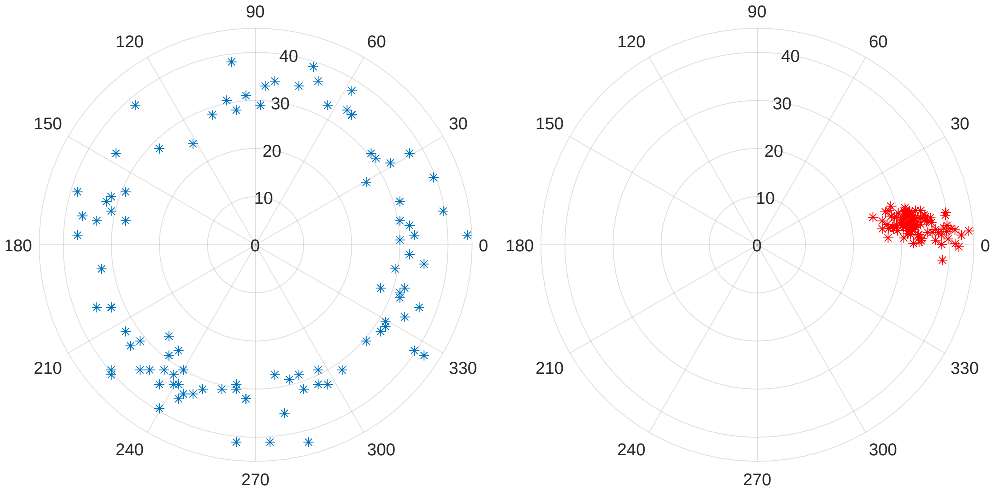
$$a = \frac{\angle\widehat{H}(f)_K - \angle\widehat{H}(f)_1}{m_K - m_1}, \tag{2.27}$$

$$b = \frac{1}{K}\sum_{k=1}^{K} \angle\widehat{H}(f)_\kappa. \tag{2.28}$$

Afterwards, the calibrated phases $\angle\tilde{H}(f)_\kappa$ for the k-th subcarrier, as shown in Fig. 2.10(b), are computed as follows:

$$\angle\tilde{H}(f)_\kappa = \angle\widehat{H}(f)_\kappa - am_\kappa - b. \tag{2.29}$$

Subtracting the linear term $am_k + b$ from the raw measurement $\angle\widehat{H}(f)_\kappa$ it is possible to obtain a linear combination of genuine phases, denoted as $\angle\tilde{H}(f)_\kappa$, from which the random phase offsets have been removed.



(a)

(b)

Figure 2.10. CSI extracted phase processing example. In (a) and (b), the raw and sanitized phases of a single subcarrier, respectively.

Chapter 3

Person Re-ID through Radio Biometric Signatures

This chapter describes the deep neural network architecture designed to solve the person Re-ID task. Initially, an overview of the proposed method is presented, then the Wi-Fi signal processing, generation of radio biometric signatures, and Re-ID are introduced in detail.

3.1 Proposed Method

An architecture that expands a two-branch siamese structure, comprising two parallel sub-networks per model branch, was designed to achieve person Re-ID from Wi-Fi signals, as shown in Fig. 3.1. The whole network can exploit both signals amplitude and phase to address the re-identification task by following the proposed pipeline. The rationale behind this choice is twofold and a requirement for the person Re-ID task, where the lack of annotated data must be managed [150, 187, 101]. First, this design is efficient in both supervised and unsupervised feature space learning. Second, the underlying strategy allows to extract invariant feature representations that enable a distance-based Re-ID, where similar identities will be close in the learned feature space while dissimilar persons will have distant representations [171]. Moreover, a siamese strategy [18] has been widely used to address the vision-based person Re-ID, achieving remarkable results [51, 183]. However, its usage for other wireless sensing applications is fairly new and still being developed [164].

Specifically, the system first performs a CSI estimation step to capture propagation properties of signals influenced by humans standing between the transmitting and receiving APs. The CSI measurement containing the affected signal is then employed to extract amplitudes and phases, which are in turn processed to generate sanitized feature vectors that represent relevant radio biometrics of a given person. In particular, filtered amplitudes are transformed into heatmaps and analyzed through a CNN-based network to capture meaningful signal patterns; calibrated phases are instead processed via an LSTM-based model to describe discriminant temporal changes deriving from life processes such as respiration and heartbeat. Subsequently, the two sub-networks (i.e., CNN and LSTM components) outputs are combined into a single feature vector representing a radio biometric signature that can be

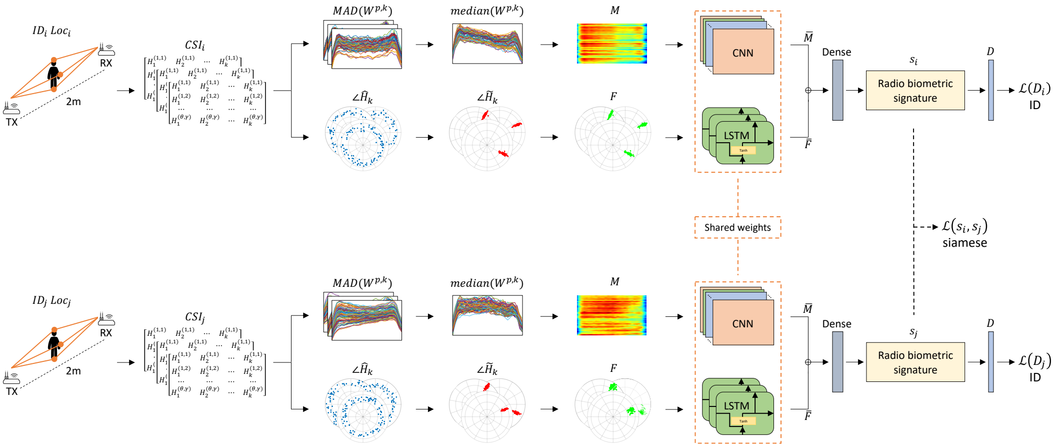


Figure 3.1. Proposed model architecture for person Re-ID through Wi-Fi. Starting from a wireless transmission, CSI is estimated and used to extract amplitude heatmaps and phase vectors as radio biometrics. A CNN and LSTM unit are then exploited to build relevant radio biometric signatures, used for the person re-identification.

used to re-identify a person across the same or at different locations. Notice that the proposed model, due to the identical branches with shared weights, is suitable for finding similarities between comparable inputs and can generate final feature vectors, i.e., radio biometric signatures, that account for possible environment noise derived, for example, from different furniture. Indeed, by following the classical vision-based siamese objective function structure, the proposed method ensures that signals associated to the same person will have similar representations in the feature space; therefore enabling for their Re-ID.

3.1.1 Channel State Information Estimation

The first CSI estimation step leverages commodity hardware for the TX and RX APs, fixed in place inside stationary environments to reduce the amount of random ambient noise. In detail, an 802.11n commercial router is used as transmitter (i.e., TX), while an Intel Wi-Fi Link 5300 (IWL5300) network interface card (NIC), connected to a Desktop PC, acts as receiver (i.e., RX). The latter was chosen since custom firmware and drivers that enable the CSI estimation were implemented in [54], as it is still rather uncommon to use commodity hardware to access CSI estimation. Furthermore, the proposed system exploits the MIMO technology to take full advantage of the multi-path propagation as a consequence of TX and RX APs integrating $\Gamma = 2$ and $\Theta = 3$ antennas, respectively. Formally, considering the described Wi-Fi signals propagation scenario, the communication channel can be modeled in the time domain as the linear time-invariant channel filter specified by Eq. (2.15) reported in Sec. 2.2. Note, however, that the CSI is a frequency-based measurement; thus, for its estimation, the fast Fourier transform (FFT) is applied on the impulse response at the receiver to obtain the corresponding CFR complex value [111]. Consequently, the APs time-invariant communication channel in the frequency domain can be linearly modeled as follows:

$$y = H(f)\,x + \omega \tag{3.1}$$

where y is the received signal vector; $H(f)$ represents the CFR at specific frequency f; x is the transmitted signal vector; and ω indicates the AWGN component. From this channel model, the OFDM technology provides a sampled CFR with a subcarrier granularity; therefore, the CSI measurement is computed by including the CFR value from each of them. Specifically, the IWL5300 component uses $K = 30$ OFDM subcarriers sampled from the 20MHz channel which contains 56 subcarriers. For each subcarrier $\kappa \in K$, the frequency response $H(f)_\kappa^{(\theta,\gamma)}$ over the receiving $\theta \in \Theta$ and transmitting $\gamma \in \Gamma$ antennas, can then be represented via the complex Eq. (2.20). The final CSI matrix computed over the frequency response of all the subcarriers, accounting for all transmitting and receiving antennas, is a 3 × 2 × 30 matrix defined exploiting the Eq. (2.21), as follows:

$$CSI = \begin{bmatrix} H(f)_1^{(1,1)} & H(f)_2^{(1,1)} & \dots & H(f)_{30}^{(1,1)} \\ H(f)_1^{(1,2)} & H(f)_2^{(1,2)} & \dots & H(f)_{30}^{(1,2)} \\ \vdots & \vdots & \vdots & \vdots \\ H(f)_1^{(3,2)} & H(f)_2^{(3,2)} & \dots & H(f)_{30}^{(3,2)} \end{bmatrix}, \tag{3.2}$$

where $H_\kappa^{(\theta,\gamma)}$ is a signed 8-bit complex number indicating the κ-th subcarrier CFR value over the $\theta \in \Theta$ and $\gamma \in \Gamma$ antennas. Observe that both amplitude and phase can be retrieved from the CSI matrix, but require further processing to be used by the proposed system as shown in Subsections 2.6.1 and 2.7.1.

3.1.2 Amplitude Sanitization and Heatmap Generation

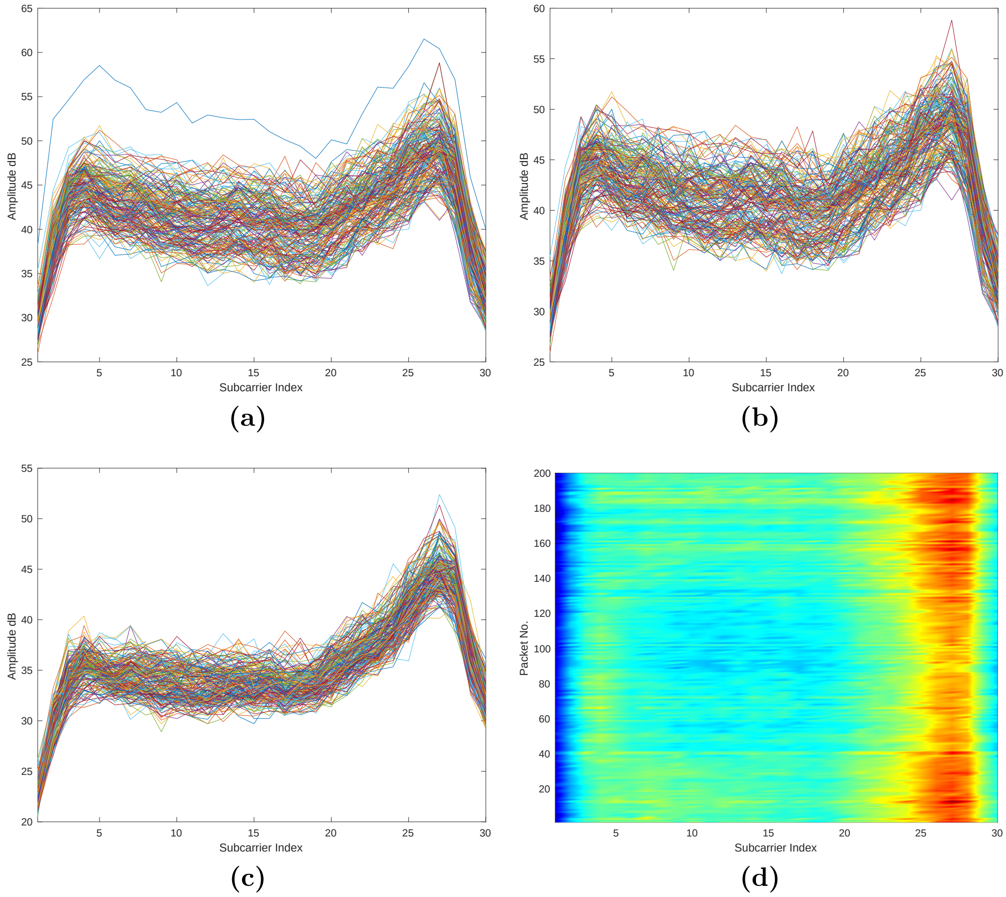
To prepare clean radio biometrics for the CNN sub-network, CSI extracted amplitudes are sanitized and transformed into heatmaps. See that the sanitization step is required since the retrieved amplitudes present noise due to various factors such as furniture material and position, external radio interference, and other environmental conditions, as shown in Fig. 3.2(a).

Concerning the sanitization procedure, the method described in Subsec. 2.6.1 is applied. Specifically, local outliers are first detected through local median values computed over a sliding window of fixed length. Subsequently, these outliers are replaced using the previous non-outlier value to retain consistent amplitude information. In particular, outliers are identified by points resulting more than $\xi = 3$ local MAD away from the local median within the sliding window applied across packets of each subcarrier. Formally, given a wireless transmission between a TX and RX antenna, amplitudes extracted from CSI measurements of $p \in P$ data packets, and a window size $w = 5$, by following the Eqs. (2.22) and (2.23) the local median is defined as:

$$\mu(\Omega^{p,\kappa}) = \Omega^{p,\kappa}_{\lceil 5/2 \rceil}, \tag{3.3}$$

$$\begin{split} \Omega^{p,\kappa} = \Big\{ |H(f)_\kappa|^{p-\lfloor 5/2 \rfloor}, \dots, |H(f)_\kappa|^{p+\lfloor 5/2 \rfloor} : \\ |H(f)_\kappa|^{p-\lfloor 5/2 \rfloor} < |H(f)_\kappa|^{p+\lfloor 5/2 \rfloor} \Big\}, \end{split} \tag{3.4}$$

where $\Omega^{p,\kappa}$ represents an ascending order set containing 5 neighboring packets amplitude $|H_\kappa|$ of the κ-th subcarrier. Note that the median described in Eq. (3.3)



(a) **(b)**

(c) **(d)**

Figure 3.2. CSI extracted amplitude processing example. In (a) and (b), the raw and sanitized amplitudes for a single TX-RX antenna pair, respectively. In (c), and (d), the median filtered amplitudes across all antenna pairs, and the corresponding heatmap used as input for the CNN sub-network.

is computed for all 3 × 2 × 30 antennas and subcarriers combinations, however a single sample is reported for the sake of simplicity. Considering Eqs. (2.24) and (2.25), the local MAD used to identify outliers and sanitize the amplitudes is then computed as follows:

$$\begin{aligned} \sigma(\Omega^{p,\kappa}) = {} & \mu(|\Omega_i^{p,\kappa} - \mu(\Omega^{p,\kappa})|), \\ & \forall i, \text{ s.t. } 1 \leq i \leq 5. \end{aligned} \tag{3.5}$$

while the acceptable local amplitude ranges are defined as:

$$limit^{p,\kappa} = \mu(\Omega^{p,\kappa}) \pm 3 * \sigma(\Omega^{p,\kappa}), \tag{3.6}$$

and every amplitude resulting outside these limits is replaced with the previous non-outlier value to maintain signal consistency. The produced sanitized signals, for an empirically chosen window size $w = 5$, are shown in Fig. 3.2(b).

Upon this first processing procedure that enables to reduce artifacts in CSI measurements, a second median filtering is applied over the 3 × 2 transmissions.

The reason behind this decision is twofold. First, it allows to reduce the data dimensionality, and second, it condensates amplitudes characteristics shared among different antennas transmissions, as shown in Fig. 3.2(c). Notice that this decision was taken since, in general, the sanitized 3×2 transmissions present similar properties. Lastly, the concentrated amplitudes are transformed into a single heatmap M of size $P \times K$, as displayed in Fig. 3.2(d), representing a person's amplitude radio biometric; which is to be used as input for the CNN sub-network.

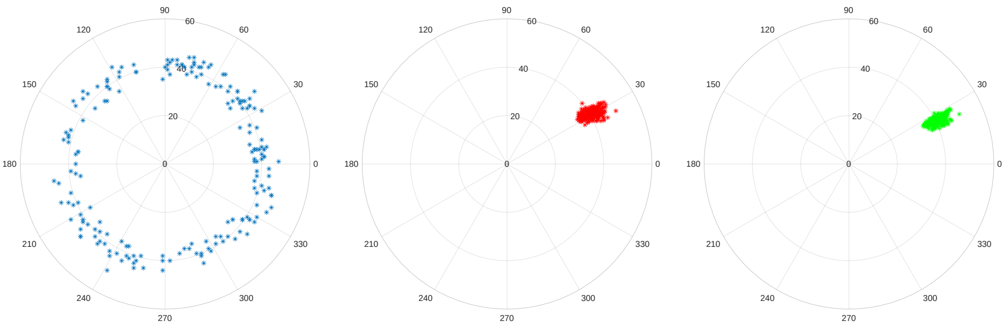
3.1.3 Phase Sanitization

Similarly to CSI amplitudes, phases also require to be processed due to common issues such as random noise and unsynchronized time clocks between TX and RX APs, that can result, among other things, in random phase offsets, as shown in Fig. 3.3(a). To address this issue, CSI-extracted are calibrated phases using the linear transformation presented in Sec. 2.7.1. Formally, a raw CSI phase $\angle\widehat{H}_\kappa$ measured for the κ-th subcarrier can be expressed as Eq. (2.26). In particular, for the IWL5300 network interface, subcarrier indices range from -15 to 15, while $N = 30$. To calibrate the phase, since the subcarrier frequency is symmetric, it possible to apply the linear transformation specified in Eq. (2.29) and ignore unknown parameters by considering the phase across the total frequency band. Specifically, the phase slope a and offset b defined by Eqs. (2.27) and (2.28), respectively, can be defined as:

$$a = \frac{\angle\widehat{H}_{30} - \angle\widehat{H}_1}{m_{30} - m_1}, \tag{3.7}$$

$$b = \frac{1}{30}\sum_{\kappa=1}^{30} \angle\widehat{H}_\kappa. \tag{3.8}$$

Afterwards, the calibrated phases $\angle\widetilde{H}_\kappa$, shown in Fig. 3.3(b), are computed by Eq. (2.29). Once the phase calibration is completed, a median filtering is applied over the 3×2 transmissions, similarly to the amplitude procedure, to reduce data dimensionality and agglomerate typical phase values along the various subcarriers, into a vector F. An example of filtered phases is shown in Fig. 3.3(c). Finally,



(a) (b) (c)

Figure 3.3. CSI extracted phase processing example. In (a), (b), and (c), the raw, sanitized and median filtered phases of a single subcarrier, respectively.

the vector F of size $P \times K$, containing the processed phases that capture temporal changes of a signal propagation, is used as input for the LSTM sub-network.

3.1.4 Radio Biometric Signatures

To perform person Re-ID an architecture based on a siamese structure is proposed; it is implemented via a two-branch neural network with parallel sub-networks in each branch, that is trained as a feature extractor, and that can learn invariant mappings [52] from the extracted radio-based features, thus resulting in a good choice for the addressed task. In detail, to compute such mappings, the model is composed by two identical branches with shared weights. Moreover, both of the presented architecture branches contain two parallel sub-networks, i.e., a CNN and an LSTM model, to correctly analyze the preprocessed signals. In particular, for the CNN module we followed a VGG-16 structure without its classification component (i.e., up to and including the last max pooling operation), since this model is an effective image pattern extractor [115]. The LSTM sub-network, instead, is implemented via a single recurrent neural network (RNN) layer containing P LSTM units. Moreover, the CNN model takes as input the heatmap M presented in Sec. 2.6.1, representing biometric information derived from amplitudes, and outputs a feature map vector $\bar{M}$; while the LSTM receives as input the F vector introduced in Sec. 2.7.1, containing temporal biometric information, and outputs the feature vector $\bar{F}$. The resulting sub-network outputs, i.e., $\bar{M}$ and $\bar{F}$, are then concatenated and merged together through a dense layer to build what is defined in this work as a radio biometric signature s.

Concerning the model training, the proposed pipeline accepts as input data pairs representing signals associated to the same or different persons. Afterwards, while biometric signatures are being learned by the sub-networks of a branch, the Euclidean distance is applied across the branches resulting outputs via a siamese loss function. This procedure allows to minimize, or maximize, the generated biometric signatures distance for similar, or dissimilar, inputs, respectively. Formally, given a pair of CSI measurements (CSI_i, CSI_j) as input, biometric signatures s_i and s_j are computed by concatenating the model branches outputs and elaborating them through a dense layer as follows:

$$\begin{aligned} s_i &= w_i(\bar{M}_i \oplus \bar{F}_i) + b_i, \\ s_j &= w_j(\bar{M}_j \oplus \bar{F}_j) + b_j, \end{aligned} \tag{3.9}$$

where $\oplus$ represents the concatenation operation; while w and b indicate the dense layer weights and bias, respectively. Subsequently, the siamese loss can be defined as:

$$\underset{siamese}{\mathcal{L}(s_i, s_j)} = \begin{cases} \frac{1}{2} \left\| s_i - s_j \right\|^2, & \text{if } i = j; \\ \frac{1}{2}\ max(m - \left\| s_i - s_j \right\|, 0)^2, & \text{if } i \neq j, \end{cases} \tag{3.10}$$

where $\|\cdot\|^2$ is the Euclidean distance; while m is a margin, empirically set to 2 in this work, that helps the dissimilar signatures separation during the optimization process. What is more, notice that this architecture also enables the Re-ID of unknown people, i.e., not observed during training, since their radio biometric signatures can still be extracted and compared at test time; where likely matching identities will be associated by lower distances among signatures, in accordance with the reported siamese loss function.

3.1.5 Joint Identification and Verification

The siamese loss function is key to the signatures generation, however, on the basis of [124, 88], the training loss function is further extended by following a joint identification and verification strategy that can improve the signatures quality. In particular, at training time, each model branch will predict person identities while the siamese cost described in Sec. 3.1.4 is also globally satisfied. Formally, the biometric signature s generated by a given branch is fed to a dense layer with dimension D, i.e., the number of known persons, and the identity loss is then implemented through a categorical cross-entropy function, as follows:

$$\underset{ID}{\mathcal{L}}(D) = -\sum_{d}^{D} y_d \log\left(\frac{\exp(d)}{\sum_{d'}^{D}\exp(d')}\right), \tag{3.11}$$

where d and y_d correspond to the predicted person identity and ground truth, respectively. Subsequently, to improve Re-ID accuracy, the identification losses computed by the two model branches are also employed in the overall training objective function, as described in the following equation:

$$\mathcal{L} = \underset{siamese}{\mathcal{L}}(s_i, s_j) + \underset{ID}{\mathcal{L}}(D_i) + \underset{ID}{\mathcal{L}}(D_j). \tag{3.12}$$

Observe that this joint objective function is only used to enable the model to extract good biometric signatures from the input signals. However, at test time, the architecture is only employed as a signature extractor. As a consequence, the identification losses are ignored while the siamese one is replaced by an Euclidean distance to address the re-identification task; where lower distances between two signatures naturally indicate more likely matching identities.

Chapter 4

Human Silhouette and Skeleton Video Synthesis through Wi-Fi signals

This chapter describes the deep learning strategy designed to solve the human silhouette (or skeleton) video synthesis task. Initially, an overview of the proposed method is presented, then the Wi-Fi signal processing, cross-modality supervision, and synthesis of human dynamics are introduced in detail. Finally, implementing a privacy-conscious system, a modified network architecture is introduced for solving a small use case combining video synthesis and human activity recognition from synthesized silhouettes or skeletons.

4.1 Proposed method

A novel two-branch neural network was designed and organized on parallel branches, sharing a decoder component to map radio signals to the visual domain and synthesize human silhouette and skeleton videos from Wi-Fi signals by emulating a teacher-student relationship. In particular, the teacher supervises the student training phase, transferring vision-based information to the associated sanitized amplitudes of the observed signal. The proposed architecture is summarized in Fig. 4.1. Precisely, the teacher model is a 3D-GAN handling visual data that, after learning the low-dimensional manifold of observed videos about human silhouettes or skeletons, produces data used as the visible ground truth for amplitudes processed by the student model. Such GAN type has been chosen since it has been proven effective in synthesizing both still images[89] and videos[162]. The student is a novel hybrid autoencoder (AE) based on LSTM[12, 9] and CNN[10, 5] architectures, inspired by the domain translation devised in [196] for image-to-image synthesis and our previous experiences on training supervision [6, 11]. In detail, the latter was specifically designed to handle amplitudes recorded over time, achieved by combining LSTM and 3D-CNN architectures to generate a latent radio representation of the signal. Then, by implementing a supervision from the teacher model, the student learns the effective mapping between radio and visual domains, i.e., translating amplitudes into silhouette or skeleton videos. Due to this distinctive design, the fundamental strategy

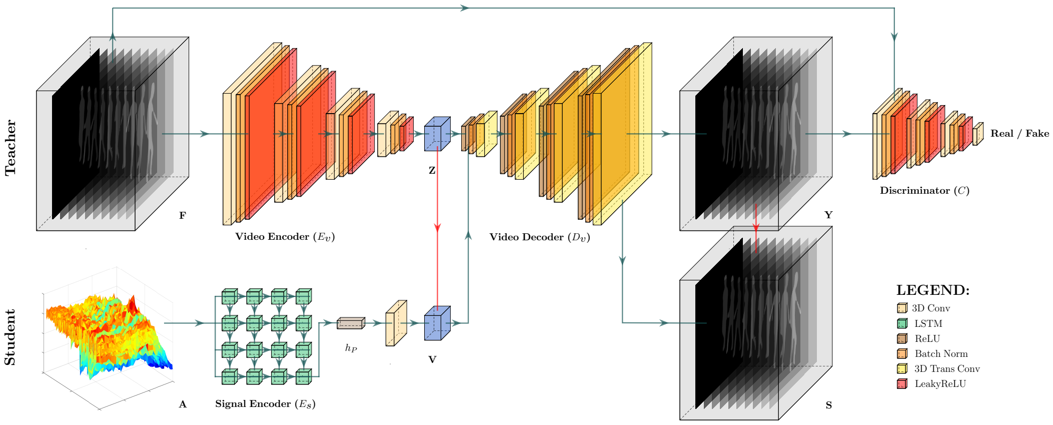


Figure 4.1. Proposed model architecture for video synthesis from Wi-Fi signals. Given a synchronized pair of human silhouette or skeleton video and CSI extracted amplitudes as input, visual knowledge is transferred to radio-based features by translating and mapping them in the visual domain via a teacher-student design. Red arrows indicate cross-modality supervision. Note that the student model can synthesize videos leveraging only Wi-Fi signals.

and difference with existing works addressing image synthesis from Wi-Fi signals, is that the proposed network architecture exploits synchronized pairs composed by a human silhouette (or skeleton) video and CSI extracted amplitudes, both taken from the same underlying environment, of a person continuously performing different poses to synthesize accurate outputs. The human silhouette, shown in Fig. 4.2, is obtained by applying the semantic image segmentation solution proposed in [22] to an input RGB sequence. The latter is also used as a starting point to extract skeletons through the OpenPose[20] framework. In this way, environment background and personal information are removed from the input, enabling the model to focus exclusively on the subject and its dynamics [4], i.e., the person moving in the scene, like in most real camera-based surveillance scenarios [8]. Instead, the sanitized amplitudes are extracted from the CSI measurements of sequential Wi-Fi data packets as signal-based features describing human poses in the radio domain [86]. This paired input enables the cross-modality supervision to learn a mapping from one domain to another during the network training phase. Accordingly, once the whole network is trained, only the student model and sanitized CSI extracted amplitudes are considered for the video synthesis. The result is a framework that can generate new person-related video frames from Wi-Fi signals without requiring any additional human or visual annotation as supervision as well as without any loss of generality. For illustration, both training and testing workflows are depicted in Fig. 4.3(a) and Fig. 4.3(b), respectively.

4.1.1 Channel State Information Estimation

Regarding a standard wireless transmission, P data packets characterize the Wi-Fi signal exchanged between fixed transmitting (TX) and receiving (RX) APs, integrating, respectively, $\Gamma > 1$ and $\Theta > 1$ antennas. Specifically, public data used for the evaluation of the proposed method were acquired with $\Gamma = 6$ and $\Theta = 3$

antennas and exploiting a 5GHz frequency band with 20MHz channel bandwidth sampling $K = 30$ OFDM subcarriers. In this MIMO setting, the CSI is measured including fine-grained signal information at the subcarrier level [166]. The CSI is a frequency-based measurement obtained by applying the fast Fourier transform (FFT) on the CIR value in time domain at the receiver, expressed as the linear time-invariant channel filter specified by Eq. (2.15) in Sec. 2.2, to compute the corresponding CFR complex number [110]. In practice, such a measurement estimates the CFR for each packet $p \in P$ reaching the RX device physical layer. Formally, in the frequency domain, the APs time-invariant communication channel is linearly modeled as follows:

$$y = H(f)\, x + \omega, \tag{4.1}$$

where y is the vector of the received signal, $H(f)$ indicates the CFR value at specific frequency f, x is the vector of the transmitted signal, and ω is the AWGN factor. From Eq. (4.1), the OFDM technology provides a sampled CFR at the subcarrier level; thus, the CSI measurement is computed by including the CFR values from K OFDM subcarriers defining the communication channel between RX and TX APs. Indeed, over the receiving $\theta \in \Theta$ and transmitting $\gamma \in \Gamma$ antennas, for each subcarrier $\kappa \in K$, the frequency response $H_{\kappa}^{(\theta,\gamma)}$ is a complex value specified by Eq. (2.20), including the signal amplitude and phase. Finally, the CSI matrix obtained accounting all communicating antennas and all subcarriers is the 6 × 3 × 30 matrix defined as:

$$CSI = \begin{bmatrix} H_1^{(1,1)} & H_2^{(1,1)} & \dots & H_{30}^{(1,1)} \\ H_1^{(1,2)} & H_2^{(1,2)} & \dots & H_{30}^{(1,2)} \\ \vdots & \vdots & \vdots & \vdots \\ H_1^{(6,3)} & H_2^{(6,3)} & \dots & H_{30}^{(6,3)} \end{bmatrix}, \tag{4.2}$$

where $H_{\kappa}^{(\theta,\gamma)}$ is the signed 8-bit complex CFR number for the κ-th subcarrier over the $\theta \in \Theta$ and $\gamma \in \Gamma$ antennas. According to the CSI specification, the amplitude can be derived from such a matrix but eventually requires further processing to be useful for Wi-Fi sensing applications as specified in Sec. 2.6.1.

4.1.2 Signal Amplitude Sanitization

To obtain meaningful radio-based features for synthesizing human silhouette or skeleton videos, the procedure described in Subsec. 2.6.1 is applied to filter the CSI extracted amplitudes and mitigate noises due to wireless protocol specifications and environmental conditions, as can be observed in Fig. 4.4(a). Indeed, abnormal values can appear in the CSI measurement and affect the extraction of human dynamics; therefore, such outliers should be removed [99]. Filtering the amplitude allows to suppress irrelevant radio information not necessarily correlated to human activity, i.e., mitigates noise caused by various factors such as furniture material and position or other external radio interference. To this end, the outlier value is identified exploiting the local median values computed over a sliding window of fixed length and replaced with the previous non-outlier value to keep congruous amplitude information. In detail, by setting $\xi = 3$, outliers are identified by points resulting in more than three local MAD away from the local median within the

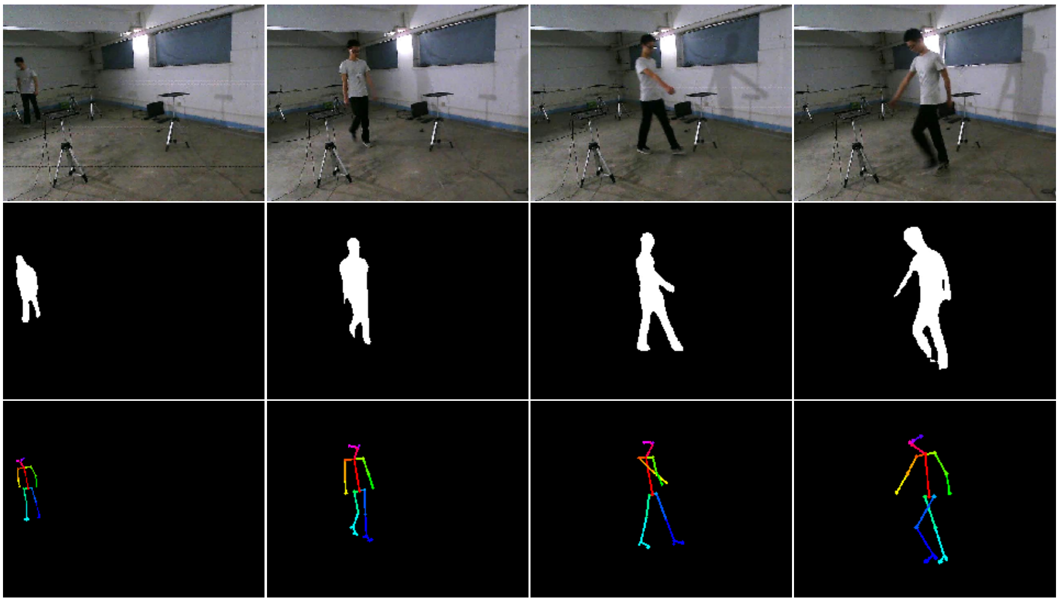


Figure 4.2. Video data input examples. The original RGB frames are reported in the top row, while the corresponding extracted human silhouette and skeleton, are shown in the middle and bottom rows, respectively.

sliding window over the packets for each subcarrier. Formally, considering the size of a window empirically set to $w = 50$, and given the signal amplitudes extracted from the CSI measurements of $p \in P$ wireless packets transmitted between the TX and RX antennas, according to the Eqs. (2.22) and (2.23), the local median is specified as follows:

$$\mu(\Omega^{p,\kappa}) = \Omega^{p,\kappa}_{\lceil 50/2 \rceil}, \tag{4.3}$$

$$\begin{aligned}\Omega^{p,\kappa} = \Big\{ &|H(f)_\kappa|^{p-\lfloor 50/2 \rfloor}, \dots, |H(f)_\kappa|^{p+\lfloor 50/2 \rfloor} : \\ &|H(f)_\kappa|^{p-\lfloor 50/2 \rfloor} < |H(f)_\kappa|^{p+\lfloor 50/2 \rfloor} \Big\},\end{aligned} \tag{4.4}$$

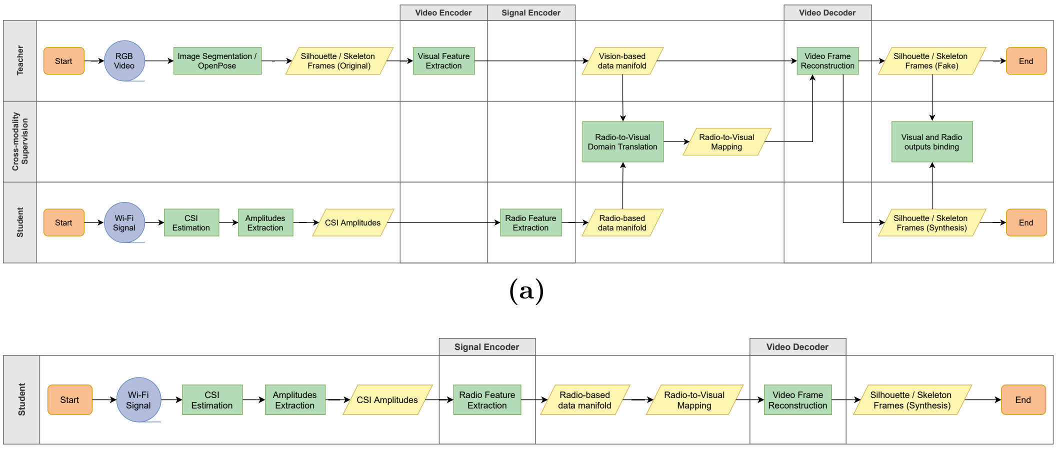
where $\Omega^{p,\kappa}$ is the set containing 50 neighboring packets amplitude $|H(f)_\kappa|$ of the κ-th subcarrier, in ascending order. Notice that the equation of a single sample is reported for the sake of simplicity; however, Eq. (4.3) is computed over all $6 \times 3 \times 30$ antennas and subcarriers combinations. Therefore, according to Eqs. (2.24) and (2.25), the local MAD used to detect abnormal amplitude values is defined as:

$$\begin{aligned}\sigma(\Omega^{p,\kappa}) = \ &\mu(|\Omega^{p,\kappa}_i - \mu(\Omega^{p,\kappa})|), \\ &\forall i, \text{ s.t. } 1 \leq i \leq 50.\end{aligned} \tag{4.5}$$

Finally, the intervals in which points are acceptable local values are defined as:

$$limit^{p,\kappa} = \mu(\Omega^{p,\kappa}) \pm 3 * \sigma(\Omega^{p,\kappa}), \tag{4.6}$$

and each amplitude falling outside these ranges is replaced with the previous non-outlier value to obtain information consistency. The produced sanitized amplitudes, for the considered window size, are shown in Fig. 4.4(b). Since the sanitized 6×3


Video Encoder
Signal Encoder
Video Decoder
Teacher
Start
RGB Video
Image Segmentation / OpenPose
Silhouette / Skeleton Frames (Original)
Visual Feature Extraction
Vision-based data manifold
Video Frame Reconstruction
Silhouette / Skeleton Frames (Fake)
End
Cross-modality Supervision
Radio-to-Visual Domain Translation
Radio-to-Visual Mapping
Visual and Radio outputs binding
Student
Start
Wi-Fi Signal
CSI Estimation
Amplitudes Extraction
CSI Amplitudes
Radio Feature Extraction
Radio-based data manifold
Silhouette / Skeleton Frames (Synthesis)
End

(a)

Signal Encoder
Video Decoder
Student
Start
Wi-Fi Signal
CSI Estimation
Amplitudes Extraction
CSI Amplitudes
Radio Feature Extraction
Radio-based data manifold
Radio-to-Visual Mapping
Video Frame Reconstruction
Silhouette / Skeleton Frames (Synthesis)
End

(b)

Figure 4.3. Proposed model workflows for video synthesis from Wi-Fi signals. In (a) and (b) the training and testing flowcharts, respectively.

transmissions share similar amplitudes properties, median filtering is applied over the transmissions to condensate such properties in a matrix A with size $P \times K$, as shown in Fig. 4.4(c) and Fig. 4.4(d). In this way, data dimensionality is reduced, and amplitudes shared among different antennas transmissions are concentrated. To train the two-branch network, the CSI extracted amplitudes are paired with the synchronized videos used to supervise the synthesis process.

4.1.3 Two-branch Network Architecture

Starting from the sanitized amplitudes paired with the corresponding synchronized video, the mapping between radio and visual features is achieved by training the two-branch network in a teacher-student fashion. In detail, to find this mapping, the network exploits two parallel branches sharing the same decoder component. The top branch has a 3D-GAN structure handling vision-based data and acts as the teacher model. Specifically, the latter consists of video frames encoder E_v, decoder D_v, and discriminator C components. In particular, for the E_v and C models, there are 3 3D convolutional layers, each comprising the sequence of strided 3D convolution, batch normalization [61], and leaky rectified linear unit (leakyReLU) [87] activation function. However, these two components differ in the last layer. In fact, being C a binary classifier, it applies a sigmoid activation function to the last 3D convolution output. Regarding the decoder D_v, it follows a reverse structure with respect to the encoder E_v with 3 3D transposed convolutions, rather than convolutions, and implements ReLU instead of LeakyReLU activation functions to stabilize the training process, as suggested in [100]. Moreover, after the last transposed convolution, D_v uses the hyperbolic tangent function to reconstruct video frames. Notice that strided 3D convolutions are used rather than traditional 2D-based ones, which are generally followed by a max-pooling operation, and the reason is twofold. First, the stride reduces the computational cost and dynamically learns the pooling operation, improving the entire model generalization [120]. Second, the 3D convolution effectively performs video analysis capturing both spatial and temporal information. Indeed, the 3D-GAN goal is to learn how to reproduce the observed human silhouette, or skeleton, videos distribution, leveraging a low-dimensional manifold that comprises feature maps in visual and temporal domains to keep track of human poses across the video frames. Intuitively, given a sequence of L 3D convolutional layers, each of them extracts spatial and temporal characteristics from the local neighborhood on the feature maps connected to various frames in the corresponding previous layer. Subsequently, a bias is applied, and an activation function is used on the result to generate feature maps on the current layer. The temporal dimension is caught convolving the 3D kernel on stacked contiguous video frames, allowing for the extraction of motion information from the video. Formally, for each feature map $j \in J_l$ computed in layer $l \in L$, the 3D value v at position

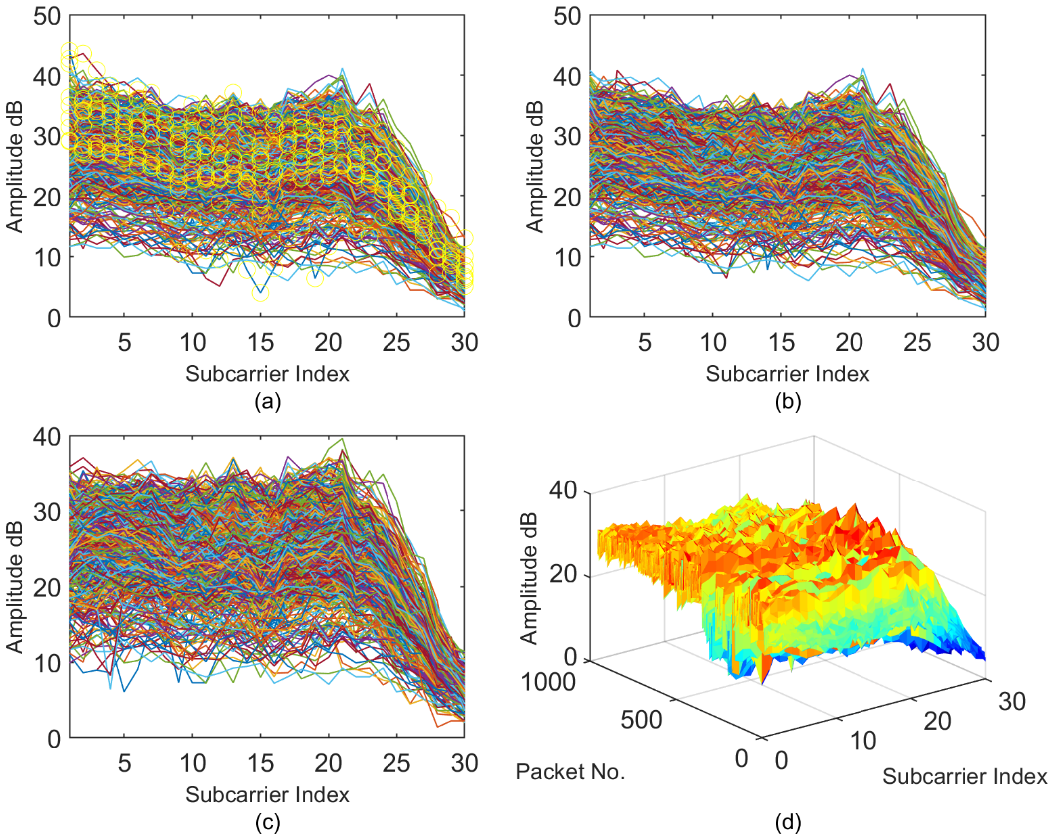


Figure 4.4. CSI extracted amplitudes processing example for 1000 data packets. In (a) and (b) the raw and sanitized amplitudes for one TX-RX antenna pair, respectively. Yellow circles are abnormal values in raw data. In (c) and (d) the median filtering over the transmissions (i.e., all antenna pairs) and the corresponding 3D surface plot for a more comprehensive view.

(x, y, z) in j is defined as:

$$v_{lj}^{xyz} = \phi\Bigg(b_{lj} + \sum_{m} \sum_{w=0}^{W_l-1} \sum_{h=0}^{H_l-1} \sum_{d=0}^{T_l-1} \rho_{ljm}^{wht} \, v_{(l-1)m}^{(x+w)(y+h)(z+t)}\Bigg), \tag{4.7}$$

where ϕ is the activation function; b_{lj} describes the bias for the current feature map; m indicates the feature map index of the previous layer $(l-1)$ connected to the j-th feature map; W_l, H_l, and T_l correspond to the height, width, and temporal depth of the 3D kernel, respectively; while ρ_{ljm}^{wht} represents the the kernel ρ_{ljm} value at position (w, h, t) connected to the m-th feature map. This characterization allows the teacher model to learn the latent space Z, with size $J_L \times T_L \times H_L \times W_L$, representing multiple contiguous frames with J_L feature maps of size $T_L \times H_L \times W_L$. Observe that this space includes low-dimensional spatial and temporal features describing human silhouette, or skeleton, poses associated to the visual domain. Concerning the bottom branch of the proposed network, it has a hybrid AE structure handling radio-based data and acts as the student model. Precisely, the latter comprises an LSTM-based signal encoder E_s and shares the video frames decoder D_v of the teacher. This type of architecture is significant for this branch because it effectively learns to map radio features to vision-based data. In particular, for the encoder E_s, there is a LSTM layer with P units, i.e., one per packet, and a 3D transposed convolution is applied on the last unit result to enable the radio-to-visual domain translation. The LSTM was chosen since it can extract features from sequential data [140]. In fact, this architecture has proven to be an ideal solution to learn the low-dimensional radio features from the sequence of CSI extracted amplitudes of contiguous wireless data packets. Afterwards, such features are translated and employed to synthesize video frames through the D_v component. Note that each LSTM unit retains important features computed from the amplitude sequence by exploiting its input, forget, and output gates to update a cell state; allowing the model to forget otherwise irrelevant information[47]. Formally, given a sequence of CSI amplitudes $A = \{a_{1,1}, a_{1,2}, \ldots, a_{P,K}\}$, over the subcarriers $\kappa \in K$, for each packet $p \in P$ the corresponding LSTM$_p$ unit is defined as:

$$\begin{aligned}
i_p &= \sigma(\Pi_{ai} a_p + U_{hi} h_{p-1} + \Psi_{ci} c_{p-1} + b_i),\\
f_p &= \sigma(\Pi_{af} a_p + U_{hf} h_{p-1} + \Psi_{cf} c_{p-1} + b_f),\\
o_p &= \sigma(\Pi_{ao} a_p + U_{ho} h_{p-1} + \Psi_{co} c_{p-1} + b_o),\\
\tilde{c}_p &= tanh(\Pi_{a\tilde{c}} a_p + U_{h\tilde{c}} \; h_{p-1} + b_{\tilde{c}}),\\
c_p &= f_p \odot c_{p-1} + i_p \odot \tilde{c}_p,\\
h_p &= o_p \odot tanh(c_p),
\end{aligned} \tag{4.8}$$

where i, f, and o indicate the the input gate, forget gate, output gate; h, $\tilde{c}$, and c denote the hidden state, cell update, and cell state, respectively; Π, U, and Ψ are the weight matrices for the corresponding gates, hidden states, and peep-hole connections; while b indicates a bias vector added to every gate or cell update. Therefore, the low-dimensional radio features learned by E_s for P packets are represented by the last LSTM unit hidden state vector h_P; capturing an abstract representation of the whole input sequence. Upon extracting these radio features, the 3D transposed convolution is applied to prepare the h_p vector for radio-to-visual domain translation, obtaining a new latent representation V that enables the silhouette, or skeleton, video frames synthesis through the D_v component.

4.1.4 Cross-modality Supervision

During the training phase, N pairs of synchronized data $< F, A >_n$ are used as the model input, where F and A correspond to the set of human silhouette, or skeleton, frames and sanitized CSI amplitudes extracted from the Wi-Fi signal associated to the video, respectively, for the n-th dataset sample. In detail, for each sample $n \in N$, the encoder E_v takes as input the set F containing real frames and computes the corresponding latent space Z. Afterwards, this low-dimensional representation is used by the decoder D_v to reconstruct the original video, defining a set of fake frames Y. On the other hand, the encoder E_s, supervised by the teacher branch, takes as input the sanitized amplitudes A and computes the latent vector h_P which, analyzed through the 3D transposed convolutional operation, outputs a latent space V, with the same shape of Z, that is used for the radio-to-visual translation. Finally, the shared component D_v synthesizes video frames S, corresponding to the original input set F, using latent space V. In general, mapping data from radio to visible spectrum is challenging due to the lack of labeled data. This problem is solved by employing the teacher branch to generate ground truth data leading to the domain-to-domain translation. In this proposed teacher-student design, the vision-based information is transferred to signal features by linking the latent representations of video and CSI extracted amplitudes.

Regarding the teacher model, the 3D-GAN learns a latent manifold of input videos required for the reconstruction goal, employing the classical GAN adversarial function based on the zero-sum game. Formally, this is achieved through the following objective loss derived from the cross-entropy between real and fake videos, defined as:

$$\begin{aligned}
&Z = E_v(F), \\
&Y = D_v(Z), \\
&\mathcal{L}_F = E_F[log\ C(F)], \\
&\mathcal{L}_Y = E_Y[log\ (1 - C(Y)], \\
&\mathcal{L}_{adv} = \min_{\{E_v, D_v\}} \max_{C} \ (\mathcal{L}_F + \mathcal{L}_Y),
\end{aligned} \tag{4.9}$$

where Z and Y are the latent space and reconstructed set of fake video frames; $C(\cdot)$ indicates the discriminator estimated probability of either real or fake videos being effectively real; while E_F and E_Y are the expected value over all the original and fake sets of frames. The proposed teacher network, generating Y from the low-dimensional features, tries to reproduce the real video F given as input. Therefore, the mean squared error is computed between video frames of F and Y, as follows:

$$MSE_Y = \frac{1}{T} \sum_{i=1}^{T} (F_i - Y_i)^2, \tag{4.10}$$

where T indicates the input video number of frames. Finally, the training objective for the 3D-GAN is computed via the following weighted equation:

$$\underset{Teacher}{\mathcal{L}} = w_{adv}\ \mathcal{L}_{adv} + w_Y\ MSE_Y, \tag{4.11}$$

where w_{adv} and w_Y, with $w_{adv} < w_Y$, are weights adjusting the impact of each objective to the overall function. Notice that the obtained latent space Z and reconstructed set of video frames Y are key elements that enable cross-modality supervision for the student model. Concerning the student model, it is implemented through a hybrid AE network that learns a low-dimensional representation of radio-based features by analyzing sanitized amplitudes via its E_s module. In addition, due to the supervision process, the student can also find a feature mapping between the visual and radio domains. Formally, given the CSI extracted amplitudes A for P contiguous wireless data packets, the latent vector h_P is computed as follows:

$$h_P = E_s(A). \tag{4.12}$$

Afterwards, a 3D transposed convolution is applied on this latent representation to obtain low-dimensional feature maps V with the same shape of Z. Subsequently, the radio-to-visual domain translation is obtained by binding the two latent spaces Z and V, effectively transferring knowledge from the visual domain, i.e., Z, into the radio one, i.e., V. Formally, this can be achieved by defining the following objective function for the student encoder E_s:

$$MSE_V = \frac{1}{|Z|}\sum_{i=1}^{|Z|}(Z_i - V_i)^2, \tag{4.13}$$

where $|Z|$ correponds to the latent space size. Moreover, to further improve the student abstraction capabilities, the reconstructed frames S, generated from the radio latent space V through the shared decoder D_v, are constrained to the set of frames Y produced by the teacher. This allows to correctly transfer knowledge by binding the two outputs, i.e., Y and S, and, consequently, to generate frames that are more similar to the real video frames F independently from the exploited input type, i.e., video or radio. Formally, this second constraint is defined as:

$$MSE_S = \frac{1}{T}\sum_{i=1}^{T}(Y_i - S_i)^2, \tag{4.14}$$

where T corresponds, again, to the input video number of frames. Notice that the two objective functions MSE_V and MSE_S are constraints required for the domain-to-domain translation between video and signal amplitudes pairs. In particular, latent space Z and reconstructed frames Y act as ground truth data enabling cross-modality supervision from the teacher during the whole network training process. Finally, the training objective for the hybrid AE, i.e., student model, is computed through the following equation:

$$\underset{Student}{\mathcal{L}} = w_V\ MSE_V + w_S\ MSE_S, \tag{4.15}$$

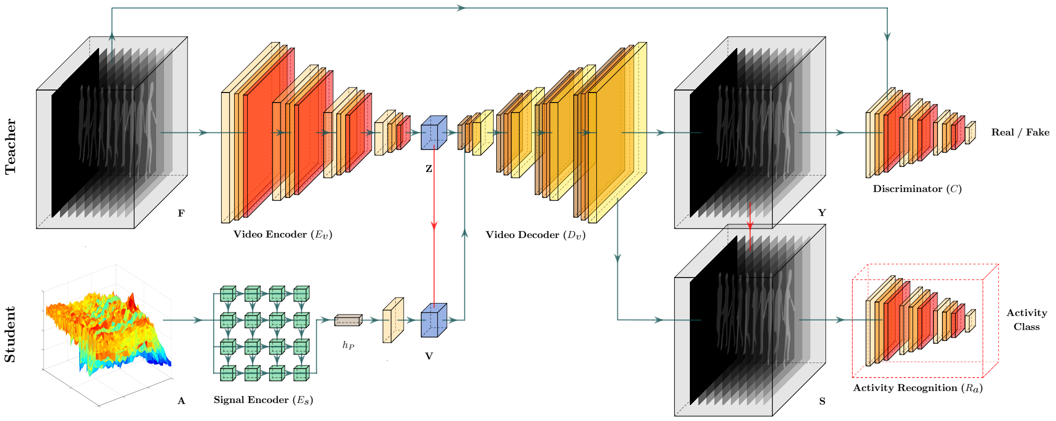


Figure 4.5. Proposed model architecture for video synthesis from Wi-Fi signals, shown in Fig. 4.1, with the addition of a human activity recognition module (i.e, the red dashed rectangle).

where w_V and w_S, with $w_V < w_S$, are weights adjusting the impact of each objective to the overall function. Concluding, the entire two-branch network objective is to minimize the following loss function:

$$\underset{Synth}{\mathcal{L}} = \underset{Teacher}{\mathcal{L}} + \underset{Student}{\mathcal{L}}. \tag{4.16}$$

4.2 Example - Use Case: Human Activity Recognition

An automated system capable of generating new visual dynamic content regarding human silhouette or skeleton from Wi-Fi signals can be an advantageous alternative or support tool in the case of limited visibility or violation of human privacy in traditional vision-based surveillance applications. Therefore, a human activity recognition module R_a was added to the network architecture designed in Fig. 4.1 to demonstrate how the proposed method works in conjunction with a simple monitoring application to understand human behavior only by analyzing silhouette or skeleton synthesized from Wi-Fi signals, satisfying the privacy-conscious concept. The new edited architecture for this specific use case is depicted in Fig. 4.5, where the new component R_a is highlighted with the red dashed rectangle. The latter has the same structure of the discriminator C except for the last layer replaced with a sequence consisting of two consecutive dense layers with leakyRelu activation function only applied on the first. In particular, the module for activity recognition will predict the human actions from synthesized videos without collecting personal data while the synthesis cost specified by Eq. (4.16) is globally satisfied. Formally, the features extracted by R_a from each generated video are fed to the second dense layer with dimension B, i.e., the number of action classes, and the action loss is then implemented through a categorical cross-entropy function, as follows:

$$\underset{activity}{\mathcal{L}} = -\sum_{b}^{B} y_b \log\left(\frac{\exp(b)}{\sum_{b'}^{B}\exp(b')}\right), \tag{4.17}$$

where b and y_b correspond to the predicted human action and ground truth, respectively. Subsequently, to effectively increase the presented model capability handling video synthesis and activity recognition jointly, obtaining a multi-task learning scheme, the overall network objective to minimize for the specified use case becomes the following loss function:

$$\mathcal{L} = \underset{Synth}{\mathcal{L}} + \underset{Activity}{\mathcal{L}}. \tag{4.18}$$

Chapter 5

Experimental Results

This chapter reports the implementation details and experimental results for both Wi-Fi sensing applications presented and discussed. For person Re-ID, the experiments are performed on a specific collected dataset fitting the task. For video synthesis, the method is evaluated employing public available Wi-Fi data focused on human subjects performing continuous poses. In addition, for the latter, the results of a use case example involving human activity recognition from synthesized data are reported to prove the usefulness of implementing the presented method in surveillance-based applications.

5.1 Person Re-ID through Radio Biometric Signatures

To present a comprehensive evaluation of the proposed methodology for person Re-ID, this section describes the data collection procedure, necessary due to the lack of public datasets for Wi-Fi Re-ID; relevant implementation details, including the chosen testing protocol and metrics; as well as qualitative and quantitative evaluations for the various system components.

5.1.1 Dataset

To compensate for the Wi-Fi Re-ID datasets unavailability, a collection was acquired to assess the presented approach. Specifically, Wi-Fi signals were captured for 35 distinct people, comprising 15 women and 20 men with similar body characteristics, standing between the TX and RX APs, for a total of 525 transmissions. In more detail, the average women height and weight were 165.3 ± 4.6cm and 61 ± 7kg, while the average men measurements corresponded to 176.1 ± 6.7cm and 76 ± 8kg. Furthermore, for each identity, five 3-seconds long transmissions (i.e., spanning over 300 packets) were collected using the 20 MHz channel of a 2.4 GHz Wi-Fi link in 3 different rooms: a conference hall, an office, and indoor hallway. Each room configuration is shown in Fig. 5.1. In all cases, the TX and RX were fixed and placed 2 metres apart, with no objects in between, and one person at a time was asked to either face toward to or away from the TX while standing between the two devices. Finally, all furniture and environment items were otherwise left untouched, and no shielding mechanism was employed to avoid interferences from other radio signals, which effectively replicates real Wi-Fi networks characteristics where multiple connections propagate across the same area and affect one another.

5.1.2 Implementation Details

Regarding the various experimental settings, Wi-Fi signals were preprocessed via the Matlab R2021a software, while several ablation studies were performed on the neural network component to correctly evaluate the proposed approach. The assessed models followed the same protocol for all tests. Specifically, the dataset was split into two subsets D_1 and D_2. The first one, which enables the model to learn how to extract meaningful signatures via the two siamese branches, contained 20 distinct people, for a total of 300 Wi-Fi transmissions. This collection was used in conjunction with a 10-fold cross-validation procedure using 4/1 random splits per person samples of each room (i.e., 240 and 60 transmissions) for the training and test sets, respectively. The second subset comprises, instead, the remaining 15 identities, counting a total of 225 samples, which were left out to evaluate the system on the re-identification of unknown people. Furthermore, for each fold, every architecture was trained for 200 epochs using the SGD algorithm [70], with an initial learning rate lr set to 0.1, a weight decay of 5e-4, and a Nesterov momentum [125] of 0.9. Moreover, a scheduler was also implemented to divide the lr by 5 at epochs 60, 120, and 160, so that the gradient update speed would be gradually reduced

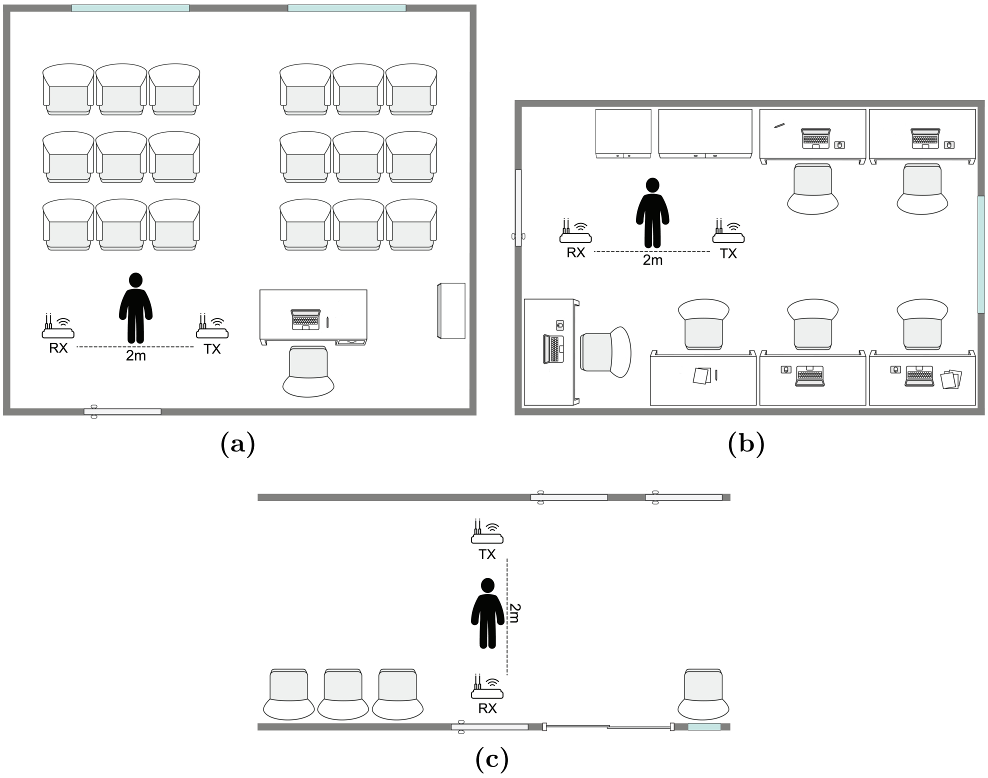
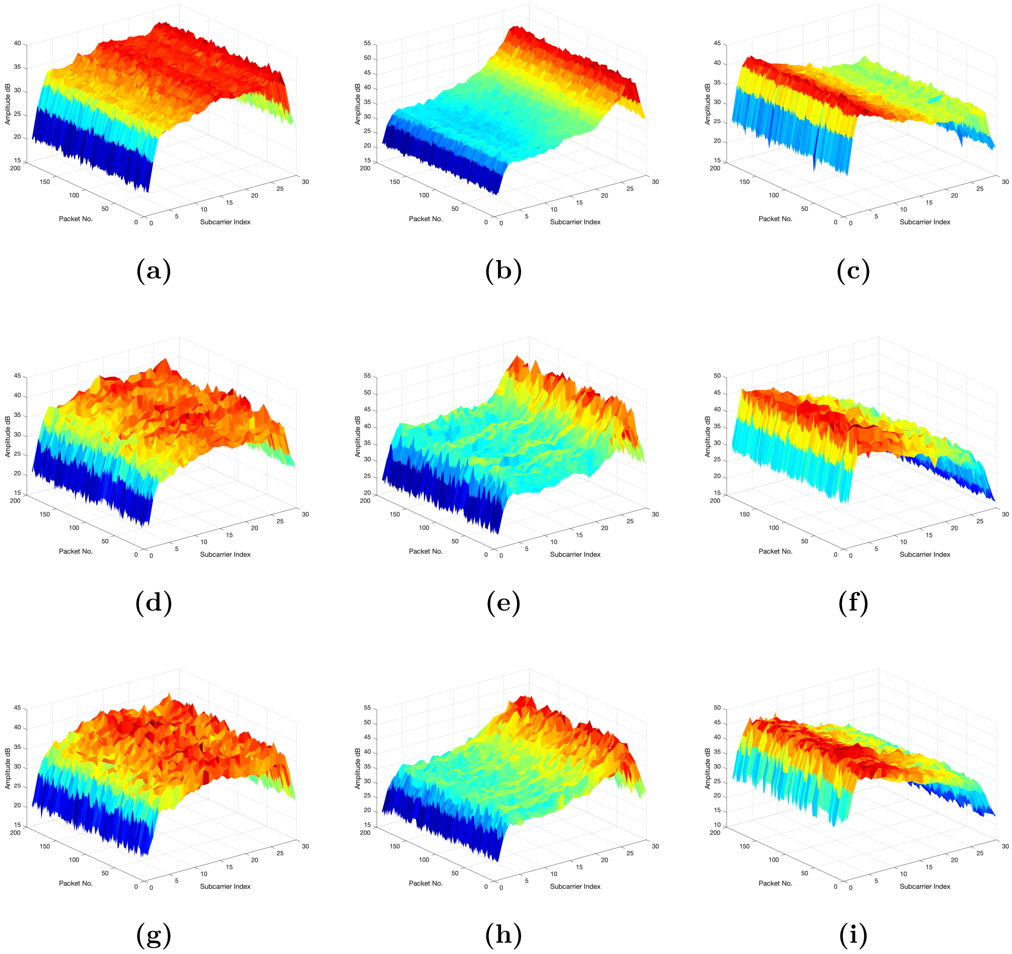


Figure 5.1. Rooms configuration for the proposed dataset acquisition protocol. In (a), (b), and (c) the conference hall, office, and indoor hallway, respectively.



(a) (b) (c)

(d) (e) (f)

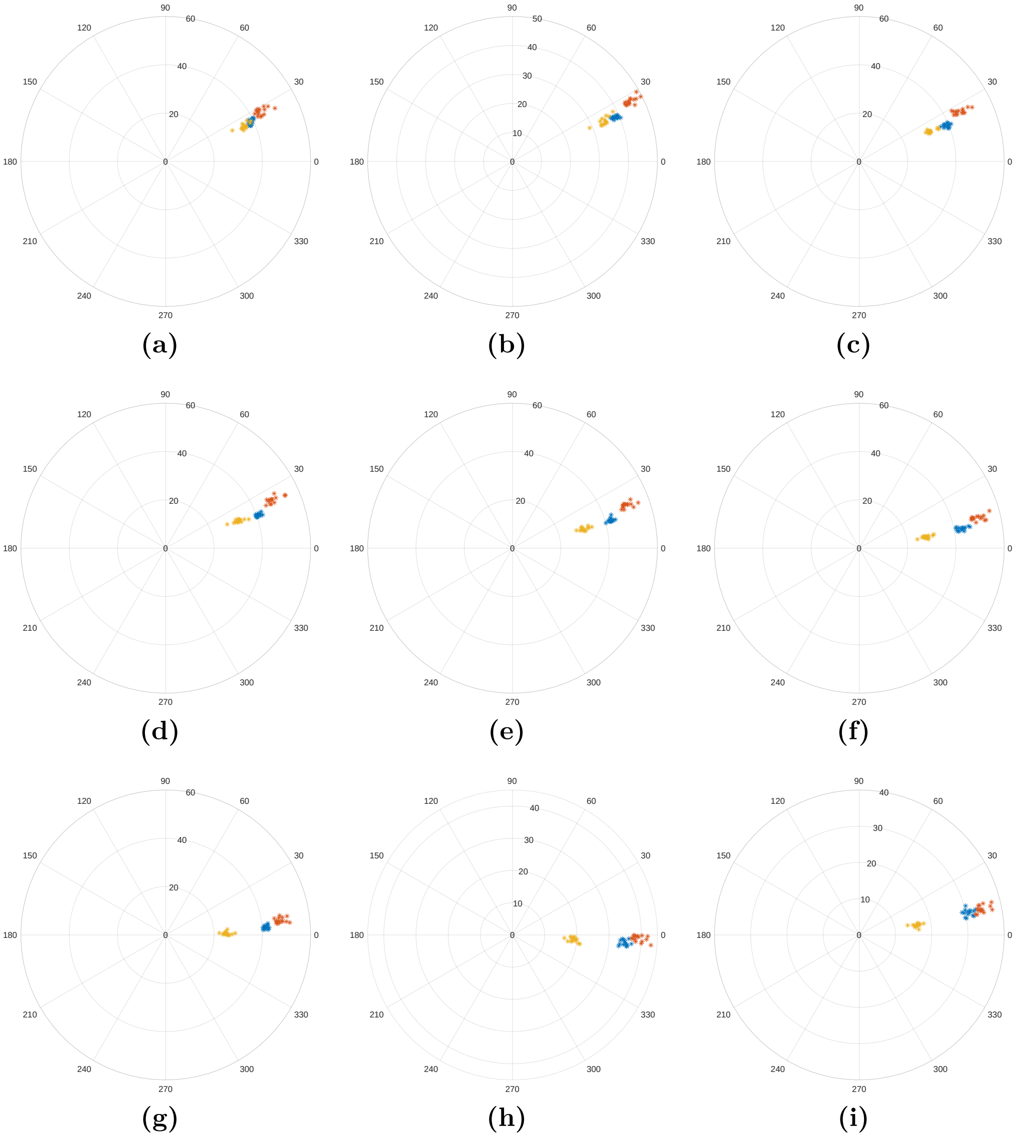
(g) (h) (i)

Figure 5.2. Heatmaps M-derived 3D surfaces examples for 200-packets acquisitions in the three rooms, i.e., a hallway, office, and conference hall, in the first, second and third row, respectively. Images (a), (d), and (g), correspond to no obstacles between TX and RX APs; while triples (b), (e), (h) and (c), (f), (i) show two different persons of the collected dataset.

for more stable signatures updates. Notice that for all experiments common person Re-ID metrics were used, such as the mean average precision (mAP) and cumulative matching characteristic (CMC) curve to represent up to Rank #10 re-identification accuracy. Finally, all networks were implemented through the PyTorch framework and its TorchVision library, while tests were performed using a single GPU, i.e., a GeForce GTX 1070 with 8GB of RAM.

5.1.3 Signals Pre-Processing Qualitative Evaluation

Amplitude and phase extracted from CSI measurements of Wi-Fi signals contain several information that can help distinguish different persons. Examples of the resulting pre-processed signals characteristics are shown in Fig. 5.2, for the amplitudes, and in Fig. 5.3 and Fig. 5.4 for the phases.



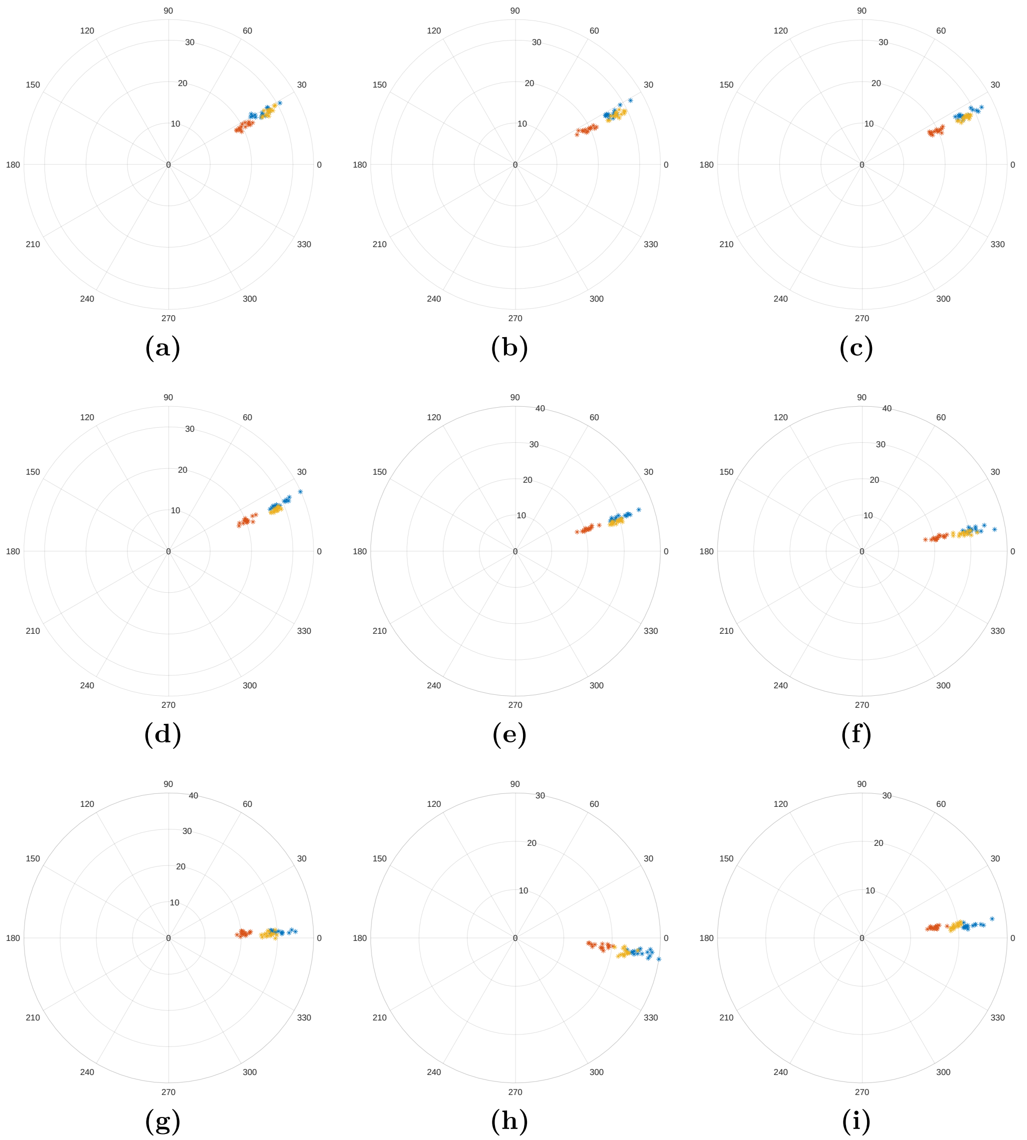
(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 5.3. Processed phases examples for 15-packets acquisitions in the same room, i.e., indoor hallway. In (a) to (i) the resulting F phases for 9 consecutive subcarriers of no obstacles between TX and RX APs and two different persons; shown in blue, red, and yellow, respectively.

Concerning the amplitudes, as can be seen in the heatmaps-derived 3D surfaces, the general shape is retained across different rooms for the same identity (i.e., first, second and third column in Fig. 5.2). However, the environment does affect the received signal even after applying the sanitizing procedure, as clearly shown in Fig. 5.2(a), Fig. 5.2(b) and Fig. 5.2(c), where the reported amplitudes are associated to the empty rooms, i.e., only furniture was present and no obstacle was left between the TX and RX APs. As a consequence, the resulting ambient noise will also affect the received amplitude quality when humans stand across the propagated signal. This outcome can be traced back to the random path followed by the signal itself,



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 5.4. Processed phases examples for 15-packets acquisitions in the same room, i.e., conference hall. In (a) to (i) the resulting F phases for 9 consecutive subcarriers of no obstacles between TX and RX APs and two different persons; shown in blue, red, and yellow, respectively.

which is not guaranteed to be the same across multiple transmissions. However, due to the stationary environments, possible random ambient noise can be successfully mitigated by operating directly in the frequency domain, which is not feasible in non-stationary scenarios, where the signal should also be processed in the time domain [149, 129]. While this result might suggest that other techniques, such as the angle of arrival, might further improve the signal processing procedures, the produced heatmaps are still able to describe human presence in a detailed way, especially when many packets are used to build the corresponding image. As a matter of fact, the heatmaps generated in the various rooms show high inter-class and low intra-class shape differences, as can be seen in each Fig. 5.2 row; therefore indicating that the derived amplitude heatmaps can correctly model characteristics of distinct persons.

Regarding the phases, they can be used to capture temporal cues from Wi-Fi signals, as can be inferred from Fig. 5.3 and Fig. 5.4. In more detail, the two images report 15 consecutive filtered sanitized phases of distinct identities for 9 adjacent subcarriers in two different rooms. i.e., indoor hallway and conference hall. As shown, as time evolves, i.e., more packets are analyzed at the RX AP, phases at each subcarrier tend to concentrate on the same spot due to the presented filtering procedure that removes phase offsets. Even more interesting, for different persons, the resulting phases have dissimilar values across the various subcarriers. This outcome suggests that first, different people also have diverse effects on the signal propagation, in accordance with the findings described in [38, 39] and second, there is a little probability for two dinstict people to have the exact distribution across all 30 subcarriers for several consecutive packets; thus indicating that a sequence-based architecture (e.g., LSTM) could most likely capture temporal shifts associated to different persons. Observe that a small number of packets is reported for each identity to avoid image clutter. However, the same reasoning applies to more subsequent packets, therefore supporting that, similarly to amplitudes, phases can also help to model unique persons and support their discrimination.

5.1.4 Wi-Fi Person Identification and Verification Evaluation

To show the effectiveness of the proposed method, several ablation studies were performed concerning the architecture, the generated signature size, as well as the number of consecutive packets to be analyzed from the Wi-Fi transmission to generate meaningful amplitude and phase features.

In relation to the chosen Re-ID model, the first batch of experiments explored the extracted features efficacy in both standalone and joint solutions by designing, respectively, a siamese architecture with single sub-network streams, elaborating either amplitudes or phases, and the presented model. Notice that these experiments were performed by using subsets with increasing complexity generated from dataset D_1. Specifically, the evaluation was performed using signals associated to either the single rooms (e.g., hallway or office), all possible pairs (e.g., hallway and office or office and conference room), and all rooms in dataset D_1 (i.e., as described in Subsec. 5.1.2). The obtained results are summarized in Table 5.1. As shown, all models achieve significant performances for both Rank #1 and mAP metrics, with the full model consistently outperforming the single-subnetwork versions (i.e., Siamese$_A$ and Siamese$_P$) by an ≈5% margin, independently of the number of examined rooms. The reason behind this outcome is twofold. First, the extracted features can capture enough differences to distinguish the 20 identities present in D_1, since each person seems to affect the signal similarly even across distinct rooms, as discussed in Subsec. 5.1.3. Second, amplitudes and phases describe different characteristics due to the chosen representation (i.e., heatmap $\bar{M}$ and temporal sequence $\bar{F}$, respectively), further improving the derived human descriptions when used jointly. Even more interesting, siamese models exploiting phase information attained lower variance across the 10-folds, which is due to temporal information captured from vector $\bar{F}$

by the LSTM unit. Indeed, while heatmaps can still represent different humans in a meaningful way, they can also be subject to higher association errors since they represent a coarse view of signals amplitudes.

Concerning the second round of ablation studies, an evaluation of different signature sizes was performed to assess the effectiveness of the fused features $\bar{M}$ and $\bar{F}$. The results obtained on dataset D_1 are reported in Table 5.2. As can be seen, employing higher dimensions for the signature s naturally results in improved performances. This is a direct consequence of the task complexity when multiple identities are present, as their representation cannot be fully described via small

Table 5.1. Model configuration 10-fold cross-validation performance evaluation on dataset D_1 for 300 consecutive packets, and $|s| = 256$. Siamese$_A$, Siamese$_P$, and Siamese models exploit amplitude, phase, and joint signal properties, respectively.

Model	#Rooms	Rank #1	mAP
Siamese$_A$	1	90.46% ± 4.40%	88.29% ± 6.36%
Siamese$_P$	1	90.12% ± 4.05%	88.17% ± 5.12%
Siamese	**1**	**94.42% ± 0.95%**	**92.90% ± 2.27%**
Siamese$_A$	2	89.78% ± 6.20%	87.96% ± 7.10%
Siamese$_P$	2	89.35% ± 4.87%	87.90% ± 5.56%
Siamese	**2**	**93.99% ± 1.01%**	**92.79% ± 2.31%**
Siamese$_A$	3	88.71% ± 7.24%	86.65% ± 7.51%
Siamese$_P$	3	88.57% ± 5.15%	86.55% ± 5.97%
Siamese	**3**	**93.51% ± 1.04%**	**92.17% ± 2.47%**

Table 5.2. Signature size 10-fold cross-validation performance evaluation on dataset D_1 for 300 consecutive packets.

Model	$\|s\|$	Rank #1	mAP
Siamese	16	56.72% ± 10.24%	50.84% ± 11.20%
Siamese	32	68.80% ± 8.05%	64.76% ± 9.06%
Siamese	64	85.59% ± 4.24%	83.41% ± 5.83%
Siamese	128	93.17% ± 1.12%	91.99% ± 2.63%
Siamese	**256**	**93.51% ± 1.04%**	**92.17% ± 2.47%**
Siamese	512	93.50% ± 1.01%	92.12% ± 2.48%
Siamese	1024	93.45% ± 0.99%	92.10% ± 2.43%

Table 5.3. Packets number 10-fold cross-validation performance evaluation on dataset D_1 with $|s| = 256$.

Model	#Packets	Rank #1	mAP
Siamese	10	80.28% ± 9.13%	79.02% ± 9.98%
Siamese	25	85.51% ± 7.99%	82.65% ± 8.65%
Siamese	50	88.50% ± 5.34%	87.23% ± 6.34%
Siamese	100	91.56% ± 3.01%	89.88% ± 4.00%
Siamese	200	93.29% ± 1.10%	92.02% ± 2.54%
Siamese	**300**	**93.51% ± 1.04%**	**92.17% ± 2.47%**

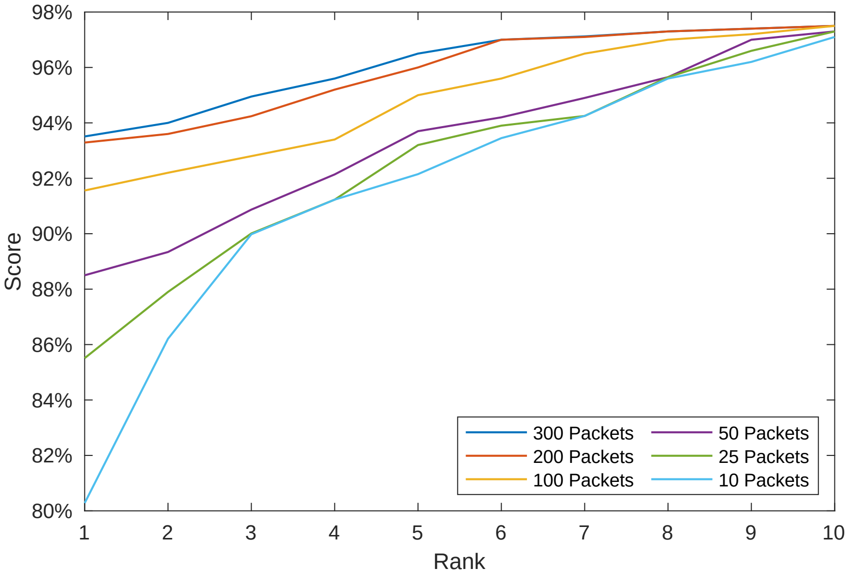


Figure 5.5. CMC curve up to Rank #10 computed on dataset D_1 for different packets number.

signatures. As a matter of fact, for $|s| < 64$, the system performances degrade rapidly and show high variance, confirming that the signature s is not able to capture meaningful characteristics for the unique identities discrimination. Moreover, there are also diminished increase returns in correspondence to bigger s sizes. This behavior is easily explained by the relatively low number of identities at our disposal (i.e., 20 for dataset D_1) which can be characterized by a signature size of 256. Nevertheless, to correctly represent as many unique persons as possible, the chosen signature size is a key component for the proposed system.

Regarding the last group of experiments, tests were performed to evaluate the effectiveness of the extracted features by modifying the number successive packets analyzed for their generation. The results obtained on dataset D_1 are summarized in Table 5.3. As can be seen, performances start converging to a stable percentage from 200 packets, indicating that the corresponding extracted features carry enough information to correctly describe the various identities in D_1. In fact, using the whole sequence of 300 packets results only in slight gains for both Rank #1 and mAP metrics. This outcome confirms the representation capability of the system, that can fully describe the various identities, while also suggesting the extracted features effectiveness. As a matter of fact, significant performances are obtained even by analyzing a lower number of packets (i.e., $P \leq 25$). However, for these configurations there is a higher variance due to the smaller extracted features which might not fully capture distinct traits for more similar radio biometrics. What is more, due to the chosen median procedures, using less than 10 packets results in a performance degradation due to an increased noise in the produced features. Nevertheless, the approach quality for fewer packets is also validated through the CMC curve shown in Fig. 5.5, where all models attain higher performances (i.e., a score of ≈90%) starting from Rank #3. This result highlights the proposed method effectiveness and its ability to represent unique radio biometric signatures which are suitable for the person re-identification task, as demonstrated in the next section.

5.1.5 Wi-Fi Person Re-Identification Evaluation

Real-world person Re-ID scenarios, such as surveillance systems, require models to also re-identify persons with different and unknown identities from those seen at training time. Therefore, to correctly evaluate the presented pipeline in such scenarios, a comprehensive assessment was performed for both model configurations and successive packets number on the distinct dataset D_2, by using the D_1-trained models with signature size $s = 256$, introduced in Section 5.1.4. Specifically, regarding the evaluation on D_2, for each of its 15 unique persons, 1 wireless transmission per room was randomly selected as the gallery, for a total of 45 transmissions; while the remaining 4 samples were used as probes to assess the re-identification capabilities of the system, counting 180 test transmissions. Moreover, since dataset D_2 represents only a small fraction of real world data, tests were performed 10 times using different random selections, and the average performance was reported to ensure statistically stable results. Concerning experiments on model configurations, the obtained results on dataset D_2 for 300 consecutive packets are reported in Table 5.4. As shown, the same behavior observed and discussed in Subsec. 5.1.4 also applies to the different unique identities of collection D_2. Indeed, by increasing the number of rooms, there is a slight performance decrease for all models, and

Table 5.4. Model configuration 10-fold cross-validation performance evaluation on dataset D_2 for 300 consecutive packets and $|s| = 256$. Siamese$_A$, Siamese$_P$, and Siamese models exploit amplitude, phase, and joint signal properties, respectively.

Model	#Rooms	Rank #1	mAP
Siamese$_A$	1	88.68% ± 3.59%	86.35% ± 4.54%
Siamese$_P$	1	88.18% ± 3.50%	86.05% ± 3.64%
Siamese	**1**	**90.28% ± 1.02%**	**89.77% ± 2.29%**
Siamese$_A$	2	86.72% ± 5.19%	80.12% ± 5.62%
Siamese$_P$	2	86.52% ± 4.24%	80.02% ± 4.78%
Siamese	**2**	**89.42% ± 1.09%**	**88.58% ± 2.55%**
Siamese$_A$	3	84.02% ± 7.25%	78.12% ± 7.70%
Siamese$_P$	3	83.62% ± 5.07%	77.62% ± 5.12%
Siamese	**3**	**88.82% ± 1.29%**	**87.52% ± 2.67%**

Table 5.5. Packets number 10-fold cross-validation performance evaluation on dataset D_2 with $|s| = 256$.

Model	#Packets	Rank #1	mAP
Siamese	10	72.12% ± 12.69%	64.52% ± 13.35%
Siamese	25	74.22% ± 9.34%	70.72% ± 10.23%
Siamese	50	80.95% ± 6.37%	78.75% ± 7.57%
Siamese	100	82.15% ± 4.98%	80.55% ± 5.60%
Siamese	200	85.05% ± 2.70%	83.85% ± 3.69%
Siamese	**300**	**88.82% ± 1.29%**	**87.52% ± 2.67%**

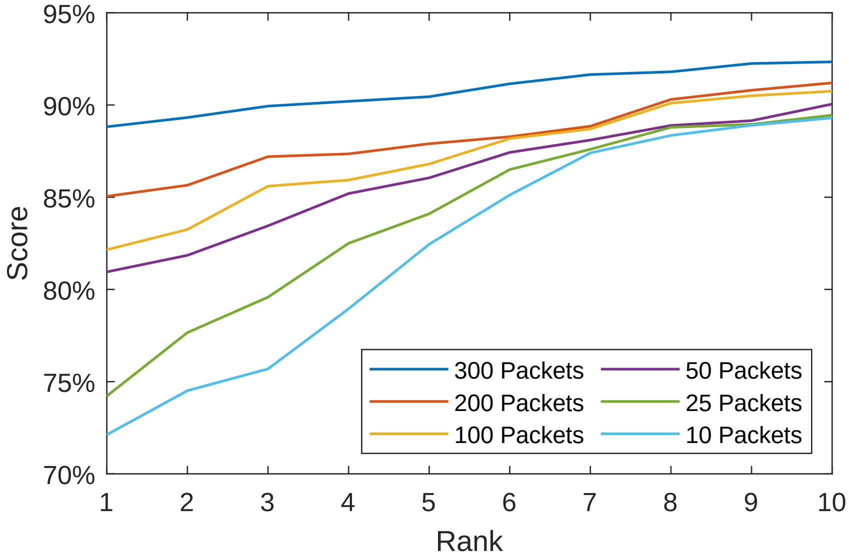


Figure 5.6. CMC curve up to Rank #10 computed on dataset D_2 for different packets number.

architectures exploiting phase information have reduced variance for both Rank #1 and mAP metrics most likely due to, as mentioned, the temporal information for the interested feature. Nevertheless, when analyzing the most complex scenario with 3 rooms, the full siamese model achieves significant perfomances; suggesting that even though the architectures have never observed the various identities, they can still extract relevant radio biometric signatures for their re-identification.

In relation to tests on the number of successive packets, results are reported in Table 5.5, while the corresponding CMC curve up to Rank #10 is depicted in Fig. 5.6. As shown, the best performing model exploits the whole sequence of 300 packets. However, differently from the performances observed in Table 5.3, where models using at least 100 packets had similar scores, for unknown identities there is a greater gap with respect to the maximum amount of recorded transmission packets. The motivation for this outcome is twofold. First and foremost, dataset D_2 has 3 times the number of test samples with respect to D_1, which was purposely built in this way to obtain consistent results over a more complex collection. Second, for real-world scenarios, i.e., where re-identification is performed on unknown people, the proposed model does not execute a training phase and, consequently, does not exploit the joint loss function shown in Eq. (3.12), which also leverages the specific identities to build more robust signatures. Nevertheless, while interesting performances are already achieved with only 10 successive packets, by increasing this number it is possible to obtain more discriminative radio biometrics and, therefore, improved radio biometric signatures able to mitigate the identity loss absence. Thus, these results confirm the findings presented in [38, 39] on signal variations in correspondence with different biological tissues, and highlight the Wi-Fi effectiveness in addressing the person re-identification task without classical vision-based drawbacks; consequently opening up a new frontier for surveillance applications where it can be crucial to re-identify unknown persons across different locations.

5.2 Human Silhouette and Skeleton Video Synthesis through Wi-Fi signals

This section first describes the public dataset used to evaluate the proposed architecture, which is focused on capturing human poses of single persons with commodity Wi-Fi. Then it provides implementation details, including the chosen hyperparameters and employed hardware. Finally, quantitative and qualitative results are reported on the public collection mentioned above to present a comprehensive evaluation of the two-branch network. Observe that, although the teacher branch is fundamental for cross-modality supervision, it is exclusively utilized during the training phase. Instead, at evaluation, the student is detached from the other branch and tested directly by using sanitized signal amplitudes as input and by comparing its reconstructed video with the real input frames paired with the Wi-Fi signals. Moreover, to further investigate the proposed cross-modality supervision strategy effectiveness in finding a mapping between radio features and vision-based human representations, experiments were performed by reconstructing either silhouette or skeleton videos exploiting exclusively radio signals as inputs.

For the image quality assessment, three state-of-the-art metrics are reported concerning the quantitative results, i.e., the mean squared error (MSE), structural similarity index (SSIM) [142], and feature similarity index (FSIM) [180]. Note that, even though the MSE is considered as the traditional measurement, it only considers pixel-by-pixel intensity comparison between original and synthesized video frames, therefore ignoring image structures. Instead, the SSIM and FSSIM address this issue by considering the structural and feature similarity, respectively. Concluding, results on a small case of study combining the silhouette and skeleton video synthesis from Wi-Fi signals to the human activity recognition are reported.

5.2.1 Dataset

The publicly available data collection presented in [50] contains 4.420 video frames in total, depicting human subjects freely performing poses in a $7m \times 8m$ room. Furthermore, each video is associated with wireless data counting 1.000 CSI samples for each RX antenna. In detail, the CSI was measured using the IWL5300 NIC implementing the CSI Tool introduced in [54], and Wi-Fi signals were acquired using 3 transceivers divided into 1 transmitter and 2 receivers, working in a 5GHz frequency band with 20MHz channel bandwidth. The former includes $\Gamma = 3$ TX antennas, while the latter each have $\Theta = 3$ RX antennas. What is more, as shown in Fig. 5.7, the receivers were placed perpendicularly to one another to increase the wireless signals resolution. As a matter of fact, according to the Fresnel zone model [147], two transceivers cannot capture a person walking parallel to the Line of Sight (LoS) path. In addition, an RGB camera was attached to a receiver to allow for synchronized video recordings and CSI samples. In particular, receivers were synchronized via the network time protocol (NTP), while wireless and video data, corresponding to a sample pair, were synchronized utilizing timestamps, with an average error of less than 1.5ms. Finally, to adapt this collection for the video synthesis task, the algorithms mentioned in Sec. 4.1 were employed to generate human silhouette and skeleton videos from the RGB sequences, i.e., semantic image segmentation and OpenPose, respectively.

5.2.2 Implementation Details

The proposed two-branch architectural design has been implemented using the Pytorch framework, and the Wi-Fi signals were processed via the MATLAB R2021a software. To correctly evaluate the proposed approach, the same protocol devised by the dataset authors in [50] is used for all the tests. Specifically, 75% of the data was used to train the network, and the remaining 25% for tests. Furthermore, each network was trained for 800 epochs using the Adam optimizer [71] with an initial learning rate set to 0.0002, an ϵ numerical stability parameter of 1e-8, first momentum term $\mathcal{B}1$ with value 0.5 and, finally, second momentum term $\mathcal{B}2$ with value 0.999. Moreover, model weights were initialized from a zero-centered Gaussian distribution with standard deviation 0.02, and the LeakyReLU slope in E_v and C components is set to 0.2. Observe that these settings are suggested for stabilizing the GAN-based networks training phase [100]. Regarding the weight parameters employed to adjust teacher and student models losses, they were empirically set to 0.5 for w_{adv} and w_V, and to 1 for w_Y and w_S. Finally, all reported experiments were performed using a single GPU, i.e., a GeForce RTX 2080 with 16GB of RAM.

5.2.3 Silhouette Synthesis Evaluation

The first batch of experiments evaluates the architecture capabilities to reconstruct human silhouette videos starting from Wi-Fi signals. The proposed solution allows the student model to learn radio-to-visual features translation which can focus exclusively on human body dynamics even though Wi-Fi signals contain coupled scattering patterns of the human body and environment. As a matter of fact, by

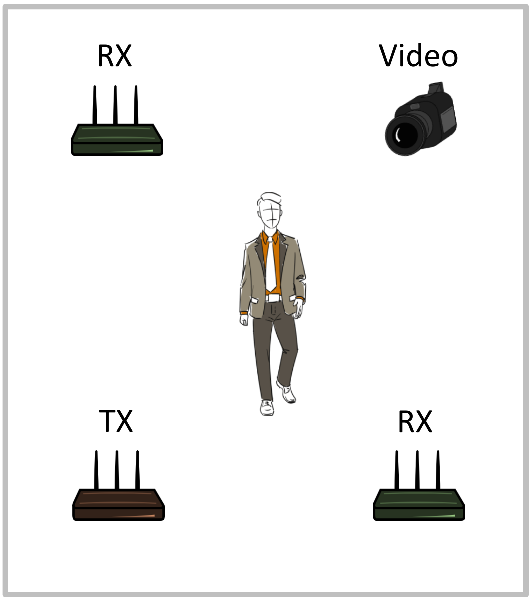


Figure 5.7. Transmitter (TX), receivers (RX), and RGB camera locations used for data collection in the experimental setting proposed in [50].

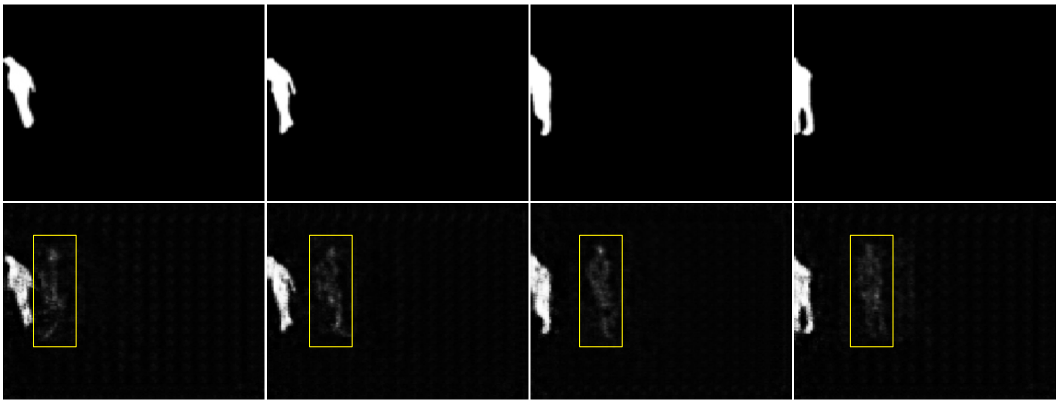


Figure 5.8. Examples of synthesized human silhouette for h_P with size of 100. In the top row, the silhouette frames representing the ground truth. In the bottom row, the noisy synthesized silhouette affected by the ghost effect, identified through a yellow rectangle.

retaining knowledge from the teacher, it is possible to transform CSI extracted amplitudes into features discriminating human-related information. More precisely, unlike existing methods that directly tune the network output, the proposed approach acts on the low-dimensional radio features latent space representation V associated with the visual domain Z via cross-modality supervision. Therefore, since V and Z have the same shape due to the architecture structure, vector h_P size results critical for the domain-to-domain translation as it regulates the amount of information extracted from radio signals. Accordingly, tests were performed to evaluate the translated latent space V effectiveness by changing the size of h_P, i.e., radio-based feature vector, ranging from 100 to 400 elements. The quantitative evaluation for this ablation study is reported in Table 5.6. As can be observed, the MSE measurement is close to zero for all h_P sizes, meaning that the student model can reconstruct accurate silhouette videos pixel-wise. The high FSIM scores also confirm the latter. In fact, this measure indicates a high image quality with respect to the expected output, i.e., ground truth silhouette videos. Significant performances are also achieved through the SSIM metric, which corresponds to structural similarities between the GT and generated output. However, by using a small h_P size, i.e., 100, the extracted signal features cannot correctly describe human movements in a scene, resulting in a noisy silhouette. This effect can be observed in Fig. 5.8, where synthesized frames present after images in the form of ghost silhouettes performing random poses due to the low representation capability derived from the small h_p size.

Table 5.6. Latent signal-based feature vector h_P size performance evaluation for human silhouette video synthesis.

h_P size	$MSE \downarrow$	$SSIM \uparrow$	$FSIM \uparrow$
100	0.007	0.781	0.972
200	0.035	0.865	0.970
300	**0.002**	**0.885**	**0.990**
400	0.002	0.828	0.984

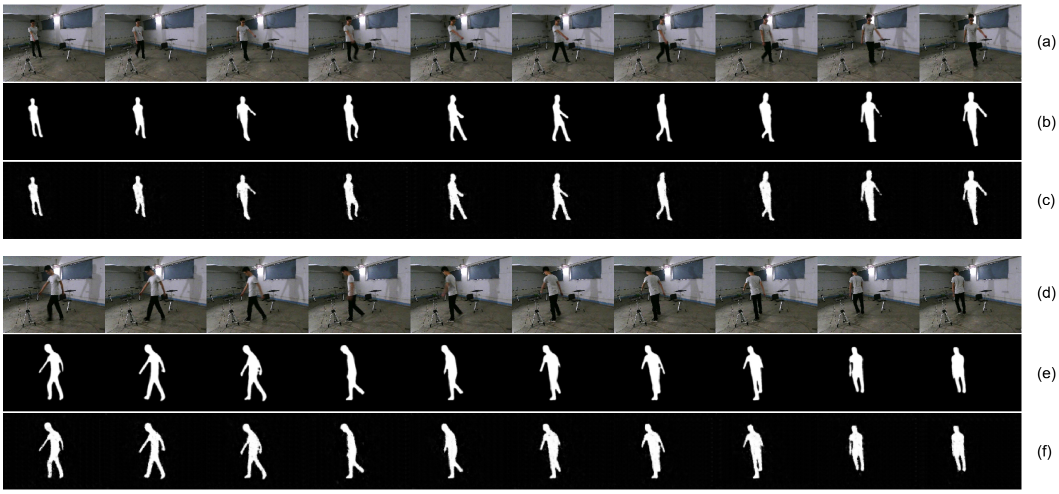


Figure 5.9. Test samples showing the human silhouette video synthesis for h_P vector with the size of 100. In (a) and (d), the original RGB video frames are reported as visual reference. In (b) and (e), silhouettes extracted from RGB frames representing the ground truth. Finally, in (c) and (f), the silhouettes synthesized exploiting exclusively Wi-Fi signals.

What is more, by increasing the radio feature size, i.e., $|h_p| = 400$, performances start to decrease due to the extracted features capturing other background information. Consequently, for human silhouette generation, the best h_p size is a vector with dimension 300. Synthesized images for this configuration are shown in Fig. 5.9, where the generated silhouettes are extremely similar and coherent with the expected output.

The human silhouette synthesis evaluation is concluded by presenting a qualitative comparison with the work introduced in [134] that, according to a thorough search of the relevant literature, is the only one performing silhouette generation from Wi-Fi signals to achieve person perception. Notice that a quantitative comparison cannot be reported since in [134] experiments are performed on a private collection and did not employ standard reconstruction metrics, such as MSE, but instead implemented a segmentation measure to evaluate their method specifically. Regardless, a qualitative comparison, albeit carried out on different images, is presented in Fig. 5.10. As can be observed, silhouettes synthesized by the presented approach, using radio features with size $|h_p| = 300$, have higher image quality and show more detailed silhouettes. Such an outcome highlights the cross-modality supervision effectiveness in this domain-to-domain translation, which is achieved by mapping Wi-Fi signals to a visual domain through the knowledge transferred from the teacher model to the student one at training time.

5.2.4 Skeleton Synthesis Evaluation

In this second group of experiments, to further assess the proposed cross-modality supervision strategy effectiveness, the architecture is evaluated by replacing silhouettes in the video-radio signal training pairs with skeleton videos as an alternative vision-based information. Human skeletons, obtained by applying the OpenPose

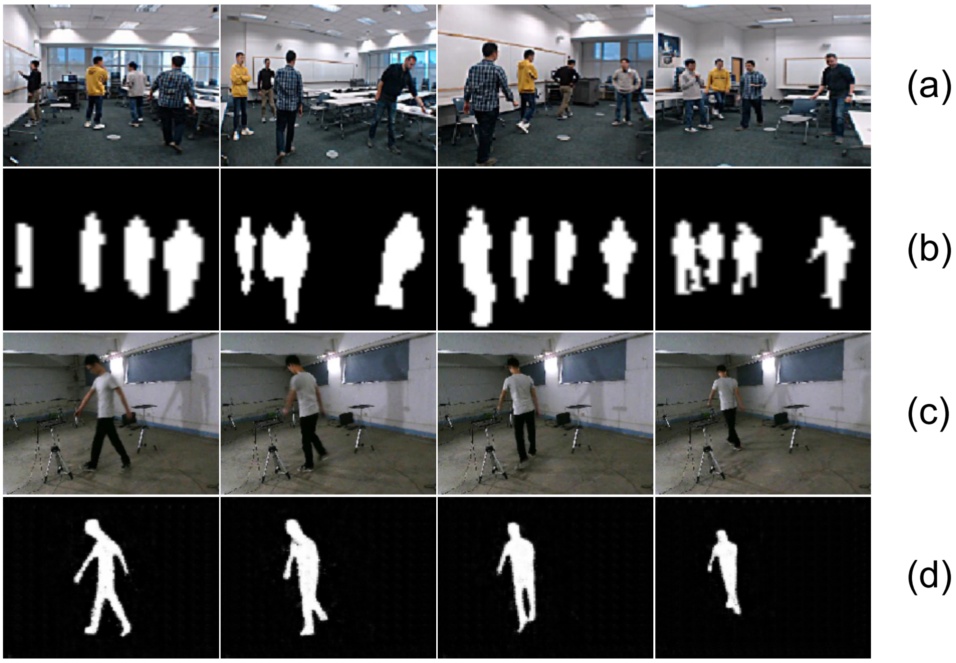


Figure 5.10. Qualitative comparison for the human silhouette synthesis. In (a) and (c), RGB visual references. In (b) and (d), the human silhouette synthesized from Wi-Fi signals in [134] and the proposed method, respectively.

framework on RGB videos, were chosen since they are one of the most widespread human body representations [7]. To train the architecture on human skeleton synthesis from Wi-Fi signals, the same implementation details described in Sec. 5.2.2 were employed. However, the system required to be trained for 1600 epochs to obtain high-quality images due to the fine-grained skeleton representation of OpenPose. Moreover, as mentioned in the previous section, since h_p regulates the amount of information used in the latent representation V used for radio-to-visual translation, the same tests on h_P vector size were performed, ranging from 100 to 400 elements. The quantitative results for the student model are summarized in Table 5.7. As shown, all h_p sizes provide roughly the same performance in feature similarity and pixel intensity comparisons, i.e., FSIM and MSE metrics, respectively. Concerning the structural similarity measure, i.e., SSIM, instead, it can be noticed that sizes 100 and 400 have higher performances in comparison with sizes 200 and 300. This outcome has a twofold explanation. First, independently from vector h_p size, enough information can be extracted from radio signals to correctly reconstruct skeleton videos. Second, the extracted radio features might be subject to noise when they have increased sizes, and it might affect the latent space representation, thus resulting in artifacts appearing inside synthesized videos, similarly to ghost silhouettes. This second aspect is caused by the OpenPose detailed skeleton, where bones connecting the various joints have different colors to indicate clearly, among other things, left from right body parts. As a matter of fact, this outcome is also supported by the higher number of epochs required for the student to reach good results on the skeleton video synthesis task. Regardless, when using a smaller h_p size, the student can avoid extracting noise, therefore synthesizing correct skeleton videos starting

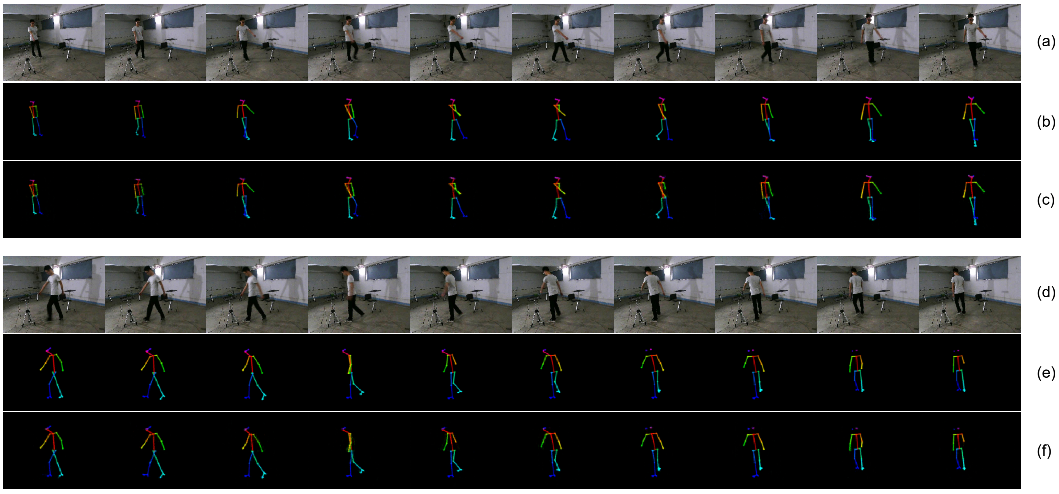


Figure 5.11. Test samples showing the human skeleton video synthesis for h_P vector with the size of 100. In (a) and (d), the original RGB video frames are reported as visual reference. In (b) and (e), skeletons extracted from RGB frames representing the ground truth. Finally, in (c) and (f), the skeletons synthesized exploiting exclusively Wi-Fi signals.

from Wi-Fi signals. The latter can be observed in Fig. 5.11, where the student model correctly reconstructs skeletons by considering a proper OpenPose color association.

In literature, due to the recent development of this field, the authors of [50] are the sole researchers currently performing experiments on the same, and only public collection available used to assess the proposed method on the skeleton image synthesis from Wi-Fi signals. Therefore, quantitative and qualitative comparisons with their work are reported to complete the human skeleton synthesis evaluation. Regarding the former, the evaluation was performed by computing the same custom metric devised in [50], i.e., percentage of correct skeleton (PCS), to have a fair comparison. The obtained results are reported in Table 5.8. In detail, the PCS metric, which is inspired by the percentage of correct keypoint (PCK), indicates the percentage of Euclidean distances between synthesized frames and their ground truths that lie within a variable threshold ξ. As shown, contrary to [50], the student model achieves remarkable performances even for minimal threshold values, indicating the synthesis of exhaustive and high-quality skeleton frames. A result that can be likely associated with the architectural design that forces the decoder component to recreate accurate skeletons from Wi-Fi signals through cross-modality supervision.

Table 5.7. Latent signal-based feature vector h_P size performance evaluation for human skeleton video synthesis.

h_P size	$MSE \downarrow$	$SSIM \uparrow$	$FSIM \uparrow$
100	**0.001**	**0.954**	**0.991**
200	0.003	0.770	0.964
300	0.005	0.776	0.949
400	0.001	0.944	0.989

Table 5.8. PCS metric comparison for different ξ values, i.e., ground truth distance. Higher percentages correspond to a better synthesis quality.

Threshold	*Student*	*Guo et al.*[50]
$\xi = 1$	95.5%	-
$\xi = 3$	96.9%	-
$\xi = 5$	98.3%	-
$\xi = 25$	100.0%	2.5%
$\xi = 30$	100.0%	26.2%
$\xi = 40$	100.0%	75.6%
$\xi = 50$	100.0%	90.0%

Regarding the qualitative comparison, synthesized skeletons are depicted in Fig. 5.12. As can be observed, even though both methods exploit OpenPose skeleton as ground truth, the presented approach synthesizes more accurate and less noisy skeletons, corroborating the results reported in Table 5.8. In fact, with respect to [50], the proposed model generates more consistent OpenPose skeletons that also take into account colors instead of binary maps, allowing to more easily identify the various limbs in the reconstructed image. Moreover, by synthesizing these detailed skeletons, the presented framework can also capture other details such as joints related to feet in the image. Such a result can be related to the extracted radio-based features that are mapped back to the visual domain by enforcing a similarity between the Z and V representations, fully highlighting and confirming the proposed cross-modality supervision and underlying architecture effectiveness.

5.2.5 Example - Use Case: Human Activity Recognition

The proposed method for silhouette or skeleton video synthesis from Wi-Fi signals enables implementing classical vision-based tasks eventually avoiding visual-related problems and protecting people privacy when personal information are not required, such as in activity recognition systems. By employing wireless signals, sensitive data are not collected; therefore, using synthesized data makes it possible to perceive

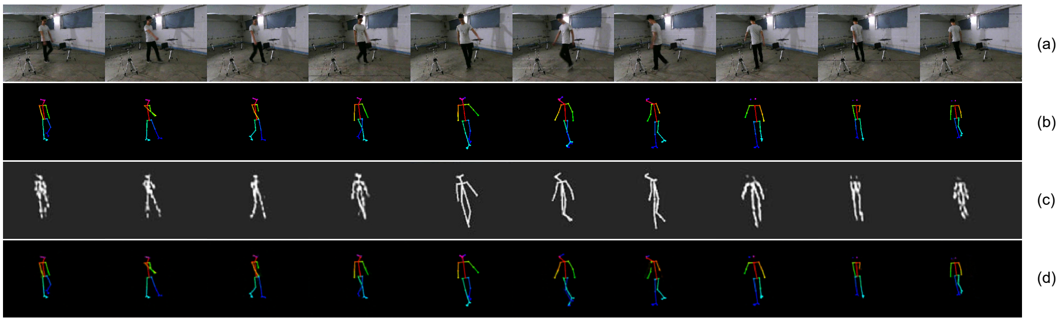
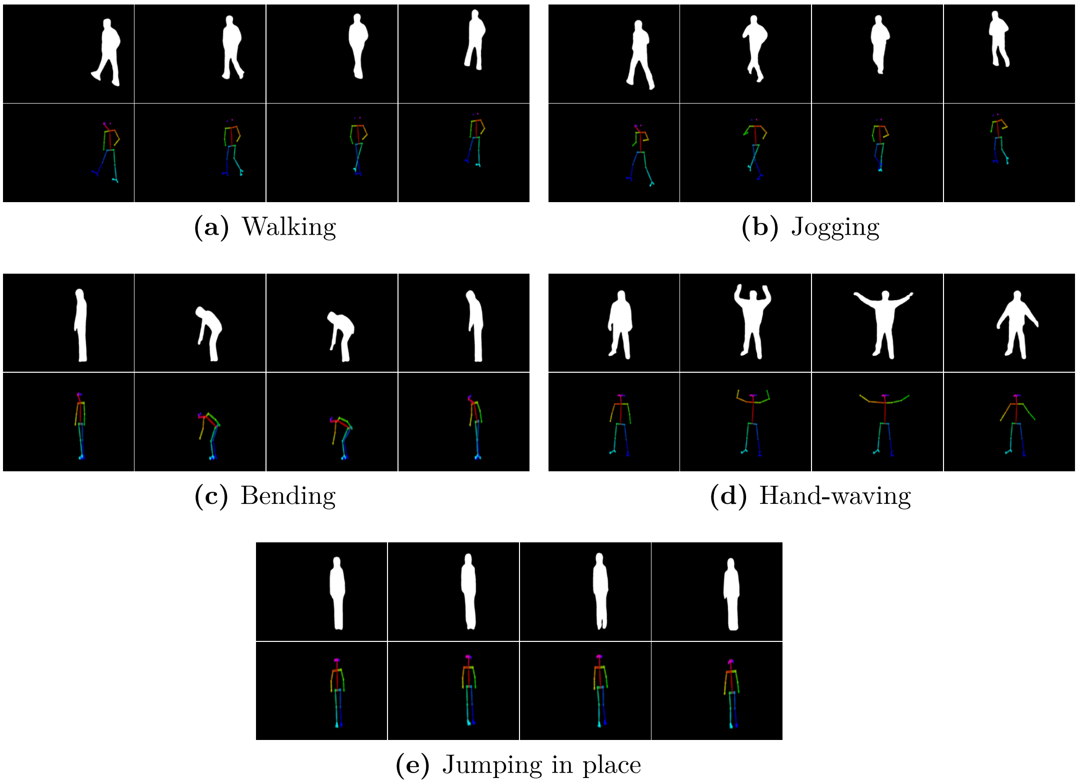


Figure 5.12. Qualitative comparison for the human skeleton synthesis. In (a) RGB frames as visual reference. In (b), the OpenPose generated ground truth. In (c) and (d), the human skeletons synthesized from Wi-Fi signals in [50] and the proposed method, respectively.

human behavior by developing an innovative and privacy-conscious system. Observing the results obtained in Subsec. 5.2.3 and Subsec. 5.2.4 on silhouette and skeleton video synthesis, respectively, the best model found per each visual data type was modified as described in Sec. 4.2 and trained to simultaneously address synthesis and activity recognition tasks as a privacy-conscious system maintaining the implementation details. To this end, a small dataset was acquired following the same acquisition protocol of public data used to evaluate the synthesis process and described in Subsec. 5.2.1. In detail, video sequences and Wi-Fi signals were captured for 5 different persons performing specific actions between the wireless transceivers: walking, jogging, hand-waving, bending, and jumping in place. The latter were chosen by selecting the common between the publicly available datasets widely used for visual activity recognition [107, 80, 96, 82]. Each person performed every action 10 different times, resulting in 50 samples in total per action. Some samples of the extracted silhouette and skeleton representations used for ground truth labeling are shown in Fig. 5.13. Due to privacy reasons, RGB video sequences are not included in this dataset. This data collection was organized using 40/10 split per person samples of each action for the training and test sets, respectively. This use case evaluation is performed reporting the classical accuracy metric and the related confusion matrix, being the activity recognition a multi-class classification problem. The classification accuracies and confusion matrices are reported in Table 5.9 and Fig. 5.14, respectively, for silhouette and skeleton synthesized data shown in Fig. 5.15. For both data type, the performance of the presented use case network structure is good; the model makes wrong predictions in a few cases due to similarities in performing some actions (e.g., walking and jogging). However, this can be solved by implementing a more advanced classifier as the R_a component of the network. The latter was not the focus of this thesis, so a simple CNN-based classifier was used. Notice that, for synthesized skeletons, the classification accuracy is barely higher, probably due to the skeleton representation that is more accurate than the human silhouette, permitting to mitigate actions similarity issues. Finally, the SSIM and FSIM metrics increased slightly for synthesized visual information, probably because the generated videos are refined through extra information, i.e., action recognition. Therefore, it was demonstrated that using Wi-Fi signals the proposed synthesis method can be used to develop human monitoring applications not revealing private or personal information useful, among the others, in many surveillance applications.

Table 5.9. Human activity recognition use case performance evaluation from both synthesized silhouette and skeleton videos.

	h_P size	$MSE \downarrow$	$SSIM \uparrow$	$FSIM \uparrow$	$ACC \uparrow$
Silhouette	300	0.002	0.952	0.995	0.880
Skeleton	100	0.001	0.983	0.993	0.940



(a) Walking

(b) Jogging

(c) Bending

(d) Hand-waving

(e) Jumping in place

Figure 5.13. Dataset samples used for ground truth labeling Wi-Fi signals and showing silhouette (i.e., top rows) and skeleton (i.e., bottom rows) performing the chosen actions.

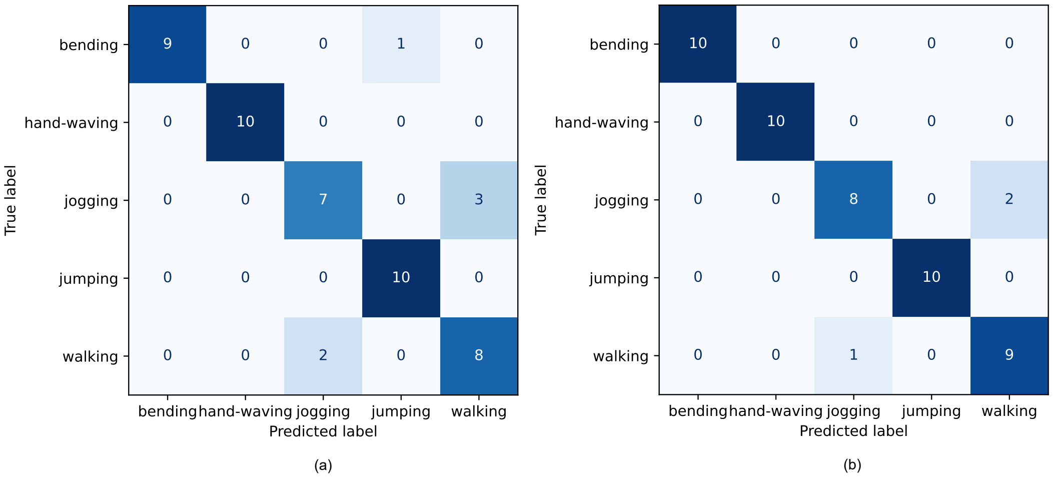
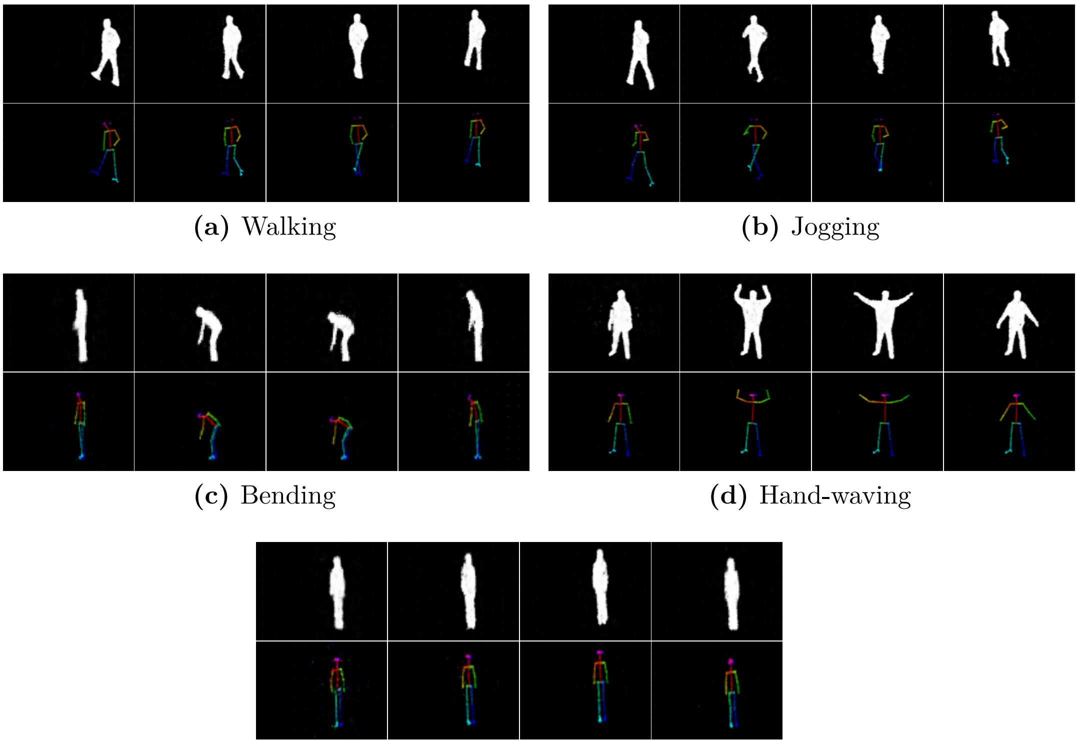


Figure 5.14. Confusion matrices for human activity recognition from (a) silhouette and (b) skeleton synthesized data.



(a) Walking

(b) Jogging

(c) Bending

(d) Hand-waving

(e) Jumping in place

Figure 5.15. Synthesized action samples from Wi-Fi signals showing silhouette (i.e., top rows) and skeleton (i.e., bottom rows) performing the chosen actions.

Chapter 6

Conclusion

This chapter concludes the thesis by first summarizing the contributions to the state-of-the-art and then by showing future work for the presented methods.

6.1 Person Re-ID through Radio Biometric Signatures

This thesis presented a novel person re-identification approach based on radio biometric signatures extracted from Wi-Fi signals. As shown by the results achieved in restricted environments, the proposed siamese architecture with parallel subnetworks, analyzing amplitude heatmaps and phase vectors, can extract meaningful representations, i.e., signatures, that enable the Re-ID of both known and unknown persons thanks to the information carried by the transmitted signals, validating the presented idea and potential application in real-world surveillance scenarios that are typically constrained.

As future work, an extended dataset version with more than 150 distinct identities will be collected and released to offer a benchmark for this unorthodox re-identification approach. In particular, this re-identification dataset will comprise multiple modalities in the form of synchronized tuples. The latter will contain Wi-Fi transmissions, RGB, and depth videos, to enable, on the one hand, a direct comparison between video and wireless modalities on the Re-ID task, and on the other hand, the implementation of multimodal methods that might benefit from the added cross-modality information when re-identifying a person across different locations. Furthermore, the presented pipeline will be used as a baseline approach for the Re-ID from Wi-Fi signals. At the same time, specific video-based and multimodal architectures will be implemented to present a comprehensive benchmark comparing the differences between the various modalities with a focus on their strengths and weaknesses. Moreover, further inquiries will be performed on other signal properties in the time domain (e.g., impulse response or time of arrival) that might be used either in a standalone solution or jointly with those implemented in this work. In addition, different solutions will also be designed to better exploit other characteristics, such as the angle of arrival, to further reduce possible ambient noise and ultimately handle more complex non-stationary environments.

6.2 Human Silhouette and Skeleton Video Synthesis through Wi-Fi signals

This thesis also introduced a novel generative Wi-Fi sensing framework capable of synthesizing human silhouette and skeleton videos by exploiting exclusively wireless signals, enabling privacy protection in people monitoring and surveillance applications. The latter was achieved by designing a cross-modality learning strategy via a two-branch network that simulates a teacher-student model. Through this configuration, the architecture can focus on human body dynamics and build a mapping between different frequency spectra, i.e., visible and radio, by being trained on synchronized video-radio signal sample pairs. Most notably, the proposed two-branch network only requires visual data inputs at training time; then, by detaching the teacher model, the student can synthesize videos starting from wireless signals inputs. Since these signals are the only source of information for frame synthesis during the model evaluation, several ablation studies were performed on the low-dimensional radio features representation transferred into the visual domain to assess both silhouette and skeleton video synthesis. The obtained results indicate that the extracted radio features can influence the domain-to-domain mapping. Moreover, qualitative comparisons with other literature works highlight the effectiveness of the devised cross-modality learning approach since it enables the student network to synthesize more accurate and less noisy silhouette and skeleton videos. Finally, the implemented use case on human activity recognition demonstrated a joint investigation, on the one hand, of transverse approaches such as multi-task learning that can further refine the generated videos through extra information, on the other hand, of person monitoring capabilities from the generated video sequences that could be employed as an additional surveillance tool in security scenarios.

As future work, a more challenging dataset will be collected to account for more elaborate human poses and more complex environments, where there is an increased signal interference and a higher number of people simultaneously present in the scene. The former will be enable to evaluate the robustness of the proposed method in real-case scenarios where radio signal absorption, deformation, and superposition are common occurrences. The latter would instead open up additional Wi-Fi sensing applications of particular interest, where multiple people could be distinguished without video devices; enabling the recognition, for instance, of group actions from Wi-Fi signals. Moreover, further investigations will be performed on different signal properties in the time domain, e.g., impulse response or time of arrival, to predict human limb coordinates other than synthesizing visual representation.

My Publications

Avola D., Bernardi M., **Cascio M.**, Cinque L., Foresti G.L., Massaroni C. **A New Descriptor for Keypoint-Based Background Modeling.** In: Image Analysis and Processing (ICIAP). Lecture Notes in Computer Science, vol 11751. Springer, 2019.
DOI: $10.1007/978-3-030-30642-7\_2$

Avola D., **Cascio M.**, Cinque L., Fagioli A., Foresti G.L., Massaroni C. **Master and Rookie Networks for Person Re-identification.** In: Computer Analysis of Images and Patterns (CAIP). Lecture Notes in Computer Science, vol 11679. Springer, 2019
DOI: $10.1007/978-3-030-29891-3\_41$

D. Avola, **M. Cascio**, L. Cinque, G. L. Foresti, C. Massaroni and E. Rodolà, **"2-D Skeleton-Based Action Recognition via Two-Branch Stacked LSTM-RNNs,"** in IEEE Transactions on Multimedia, vol. 22, no. 10, pp. 2481-2496, 2020.
DOI: $10.1109/TMM.2019.2960588$

Avola D., **Cascio M.**, Cinque L., Fagioli A., Foresti G.L., **"LieToMe: An Ensemble Approach for Deception Detection from Facial Cues,"** in International Journal of Neural Systems, vol.31, no.2, pp. 1-20, 2021.
DOI: $10.1142/S0129065720500689$

Avola D., **Cascio M.**, Cinque L., Foresti G.L., Pannone D., **"Machine learning for video event recognition,"** in Integrated Computer-Aided Engineering, vol. 28, no. 3, pp. 309-332, 2021.
DOI: $10.3233/ICA-210652$

Avola D., **Cascio M.**, Cinque L., Fagioli A., Foresti G.L., Marini M.R., Pannone D. **Analyzing EEG Data with Machine and Deep Learning: A Benchmark.** In: Image Analysis and Processing (ICIAP), vol. 13231, Springer, 2022.
DOI: $10.1007/978-3-031-06427-2_28$

Avola D., **Cascio M.**, Cinque L., Fagioli A., Foresti G.L., **"Human Silhouette and Skeleton Video Synthesis through Wi-Fi signals,"** in International Journal of Neural Systems, vol. 32, no. 5, pp. 1-20, 2022.
DOI: 10.1142/S0129065722500150

D. Avola, **M. Cascio**, L. Cinque, A. Fagioli, C. Petrioli, **"Person Re-Identification through Wi-Fi Extracted Radio Biometric Signatures,"** in IEEE Transactions on Information Foresincs and Security, vol. 17, pp. 1145-1158, 2022.
DOI: 10.1109/$TIFS$.2022.3158058

acknowledgments

First of all, I would like to thank my parents for always being present in my life, for continuous encouragement and support, allowing me to achieve important goals. A special thank goes to the March family for gifting me with the warmth of a second family.

I would like to thank my esteemed supervisors, Prof. Luigi Cinque and Prof. Chiara Petrioli, for their invaluable supervision, support, and attention during my Ph.D. degree.

I want to express my sincere gratitude to my co-advisor, Prof. Danilo Avola, for the continuous support of my Ph.D. research, his patience, and motivation. His guidance helped me in all the research and writing of this thesis. I could not have imagined having a better mentor for my Ph.D. study.

I thank my labmates, especially Alessio Fagioli and Dr. Daniele Pannone, for the stimulating discussions, the times we worked together before deadlines, and all the fun we have had in the last years.

Last but not least, I would like to thank Federico March, Ariel, Sabrina Romanò, Cristian Camuto, Gina Koutsandria, and Valerio Di Valerio, for supporting me spiritually throughout writing this thesis and my life in general.

Bibliography

[1] Adib, F. and Katabi, D. See through walls with wifi! *SIGCOMM Computer Communication Review*, **43** (2013), 75–86. doi:10.1145/2534169.2486039.

[2] Ali, K., Liu, A. X., Wang, W., and Shahzad, M. Recognizing keystrokes using wifi devices. *IEEE Journal on Selected Areas in Communications*, **35** (2017), 1175. doi:10.1109/JSAC.2017.2680998.

[3] Assayag, Y., Oliveira, H., Souto, E., Barreto, R., and Pazzi, R. Indoor positioning system using synthetic training and data fusion. *IEEE Access*, **9** (2021), 115687. doi:10.1109/ACCESS.2021.3105188.

[4] Avola, D., Bernardi, M., Cascio, M., Cinque, L., Foresti, G. L., and Massaroni, C. A new descriptor for keypoint-based background modeling. In *Image Analysis and Processing (ICIAP)*, pp. 15–25. Springer International Publishing (2019). doi:10.1007/978-3-030-30642-7_2.

[5] Avola, D., Bigdello, M. J., Cinque, L., Fagioli, A., and Marini, M. R. R-signet: Reduced space writer-independent feature learning for offline writer-dependent signature verification. *Pattern Recognition Letters*, **150** (2021), 189.

[6] Avola, D., Cascio, M., Cinque, L., Fagioli, A., Foresti, G. L., and Massaroni, C. Master and rookie networks for person re-identification. In *International Conference on Computer Analysis of Images and Patterns*, pp. 470–479 (2019). doi:10.1007/978-3-030-29891-3_41.

[7] Avola, D., Cascio, M., Cinque, L., Foresti, G. L., Massaroni, C., and Rodolà, E. 2-d skeleton-based action recognition via two-branch stacked lstm-rnns. *IEEE Transactions on Multimedia*, **22** (2020), 2481. doi:10.1109/TMM.2019.2960588.

[8] Avola, D., Cascio, M., Cinque, L., Foresti, G. L., and Pannone, D. Machine learning for video event recognition. *Integrated Computer-Aided Engineering*, **28** (2021), 309. doi:10.3233/ICA-210652.

[9] Avola, D., Cinque, L., De Marsico, M., Fagioli, A., and Foresti, G. L. Lietome: Preliminary study on hand gestures for deception detection via fisher-lstm. *Pattern Recognition Letters*, **138** (2020), 455.

[10] Avola, D., Cinque, L., Diko, A., Fagioli, A., Foresti, G. L., Mecca, A., Pannone, D., and Piciarelli, C. Ms-faster r-cnn: Multi-stream backbone for improved faster r-cnn object detection and aerial tracking from uav images. *Remote Sensing*, **13** (2021), 1670.

[11] Avola, D., Cinque, L., Fagioli, A., Filetti, S., Grani, G., and Rodolà, E. Multimodal feature fusion and knowledge-driven learning via experts consult for thyroid nodule classification. *IEEE Transactions on Circuits and Systems for Video Technology*, **early access** (2021), 1. `doi:10.1109/TCSVT.2021.3074414`.

[12] Avola, D., Cinque, L., Fagioli, A., Foresti, G. L., and Massaroni, C. Deep temporal analysis for non-acted body affect recognition. *IEEE Transactions on Affective Computing*, **early access** (2020), 1.

[13] Avola, D., Cinque, L., Fagioli, A., Foresti, G. L., Pannone, D., and Piciarelli, C. Bodyprint—a meta-feature based lstm hashing model for person re-identification. *Sensors*, **20** (2020), 5365. `doi:10.3390/s20185365`.

[14] Benito-Picazo, J., Domínguez, E., Palomo, E. J., and López-Rubio, E. Deep learning-based video surveillance system managed by low cost hardware and panoramic cameras. *Integrated Computer-Aided Engineering*, **27** (2020), 373. `doi:10.3233/ICA-200632`.

[15] Bialer, O., Raphaeli, D., and Weiss, A. J. A time-of-arrival estimation algorithm for ofdm signals in indoor multipath environments. *Signal Processing*, **169** (2020), 107375. `doi:10.1016/j.sigpro.2019.107375`.

[16] Bianchi, V., Ciampolini, P., and De Munari, I. Rssi-based indoor localization and identification for zigbee wireless sensor networks in smart homes. *IEEE Transactions on Instrumentation and Measurement*, **68** (2018), 566. `doi:10.1109/TIM.2018.2851675`.

[17] Booranawong, A., Jindapetch, N., and Saito, H. A system for detection and tracking of human movements using rssi signals. *IEEE Sensors Journal*, **18** (2018), 2531. `doi:10.1109/JSEN.2018.2795747`.

[18] Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R. Signature verification using a" siamese" time delay neural network. *Advances in Neural Information Processing Systems*, **6** (1993), 737. `doi:10.5555/2987189.2987282`.

[19] Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43** (2021), 172. `doi:10.1109/TPAMI.2019.2929257`.

[20] Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., and Sheikh, Y. A. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43** (2019), 17432.

[21] Chang, R. W. Synthesis of band-limited orthogonal signals for multichannel data transmission. *The Bell System Technical Journal*, **45** (1966), 1775. `doi:10.1002/j.1538-7305.1966.tb02435.x`.

[22] Chen, L., Papandreou, G., Schroff, F., and Adam, H. Rethinking atrous convolution for semantic image segmentation. *CoRR*, **abs/1706.05587** (2017), 1.

[23] Chen, W. and Hays, J. Sketchygan: Towards diverse and realistic sketch to image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9416–9425 (2018). `doi:10.1109/CVPR.2018.00981`.

[24] Chen, Y.-C., Zhu, X., Zheng, W.-S., and Lai, J.-H. Person re-identification by camera correlation aware feature augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **40** (2017), 392. `doi:10.1109/TPAMI.2017.2666805`.

[25] Chen, Z., Zhang, L., Jiang, C., Cao, Z., and Cui, W. Wifi csi based passive human activity recognition using attention based blstm. *IEEE Transactions on Mobile Computing*, **18** (2018), 2714. `doi:10.1109/TMC.2018.2878233`.

[26] Chung, D., Tahboub, K., and Delp, E. J. A two stream siamese convolutional neural network for person re-identification. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1983–1991 (2017). `doi:10.1109/ICCV.2017.218`.

[27] Dang, X., Si, X., Hao, Z., and Huang, Y. A novel passive indoor localization method by fusion csi amplitude and phase information. *Sensors*, **19** (2019), 875. `doi:10.3390/s19040875`.

[28] Davies, L. and Gather, U. The identification of multiple outliers. *Journal of the American Statistical Association*, **88** (1993), 782. `doi:10.1080/01621459.1993.10476339`.

[29] Denis, S., Berkvens, R., and Weyn, M. A survey on detection, tracking and identification in radio frequency-based device-free localization. *Sensors*, **19** (2019), 1. `doi:10.3390/s19235329`.

[30] Ding, J., Wang, Y., and Fu, X. Wihi: Wifi based human identity identification using deep learning. *IEEE Access*, **8** (2020), 129246. `doi:10.1109/ACCESS.2020.3009123`.

[31] Fan, L., Li, T., Fang, R., Hristov, R., Yuan, Y., and Katabi, D. Learning longterm representations for person re-identification using radio signals. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10696–10706 (2020). `doi:10.1109/CVPR42600.2020.01071`.

[32] Fang, Y., Deng, W., Du, J., and Hu, J. Identity-aware cyclegan for face photo-sketch synthesis and recognition. *Pattern Recognition*, **102** (2020), 107249. `doi:10.1016/j.patcog.2020.107249`.

[33] Feng, Z., Lai, J., and Xie, X. Learning view-specific deep networks for person re-identification. *IEEE Transactions on Image Processing*, **27** (2018), 3472. doi:10.1109/TIP.2018.2818438.

[34] Feng, Z., Lai, J., and Xie, X. Learning modality-specific representations for visible-infrared person re-identification. *IEEE Transactions on Image Processing*, **29** (2019), 579. doi:10.1109/TIP.2019.2928126.

[35] Fischer, H. *A History of the Central Limit Theorem: From Classical to Modern Probability Theory*. Springer (2011). ISBN 9780387878560. doi:10.1007/978-0-387-87857-7.

[36] Friis, H. A note on a simple transmission formula. *Proceedings of the IRE*, **34** (1946), 254. doi:10.1109/JRPROC.1946.234568.

[37] Fu, Y., Chen, P., Yang, S., and Tang, J. An indoor localization algorithm based on continuous feature scaling and outlier deleting. *IEEE Internet of Things Journal*, **5** (2018), 1108. doi:10.1109/JIOT.2018.2795615.

[38] Gabriel, S., Lau, R., and Gabriel, C. The dielectric properties of biological tissues: Ii. measurements in the frequency range 10 hz to 20 ghz. *Physics in medicine & biology*, **41** (1996), 2251. doi:10.1088/0031-9155/41/11/002.

[39] Gabriel, S., Lau, R., and Gabriel, C. The dielectric properties of biological tissues: Iii. parametric models for the dielectric spectrum of tissues. *Physics in medicine & biology*, **41** (1996), 2271. doi:10.1088/0031-9155/41/11/003.

[40] Galama, Y. and Mensink, T. Itergans: Iterative gans to learn and control 3d object transformation. *Computer Vision and Image Understanding*, **189** (2019), 102803. doi:10.1016/j.cviu.2019.102803.

[41] Gao, Q., Wang, J., Ma, X., Feng, X., and Wang, H. Csi-based device-free wireless localization and activity recognition using radio image features. *IEEE Transactions on Vehicular Technology*, **66** (2017), 10346. doi:10.1109/TVT.2017.2737553.

[42] Ghany, A. A., Uguen, B., and Lemur, D. A robustness comparison of measured narrowband csi vs rssi for iot localization. In *IEEE Veh. Technol. Conf.*, pp. 1–5 (2020). doi:10.1109/VTC2020-Fall49728.2020.9348854.

[43] Giordano, N. *College Physics: Reasoning and Relationships*. Cengage Learning (2009). ISBN 9780534424718.

[44] Gong, L., Yang, W., Zhou, Z., Man, D., Cai, H., Zhou, X., and Yang, Z. An adaptive wireless passive human detection via fine-grained physical layer information. *Ad Hoc Networks*, **38** (2016), 38. doi:https://doi.org/10.1016/j.adhoc.2015.09.005.

[45] Gong, N., Yang, Y., Liu, Y., and Liu, D. Dynamic facial expression synthesis driven by deformable semantic parts. In *International Conference on Pattern Recognition (ICPR)*, pp. 2929–2934 (2018). doi:10.1109/ICPR.2018.8545831.

[46] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, vol. 27, pp. 2672–2680 (2014).

[47] Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., and Schmidhuber, J. Lstm: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, **28** (2017), 2222. doi:10.1109/TNNLS.2016.2582924.

[48] Gu, Y., Ren, F., and Li, J. Paws: Passive human activity recognition based on wifi ambient signals. *IEEE Internet of Things Journal*, **3** (2015), 796. doi:10.1109/JIOT.2015.2511805.

[49] Guo, L., Lu, Z., Wen, X., Zhou, S., and Han, Z. From signal to image: Capturing fine-grained human poses with commodity wi-fi. *IEEE Communications Letters*, **24** (2019), 802. doi:10.1109/LCOMM.2019.2961890.

[50] Guo, L., Lu, Z., Wen, X., Zhou, S., and Han, Z. From signal to image: Capturing fine-grained human poses with commodity wi-fi. *IEEE Communications Letters*, **24** (2020), 802. doi:10.1109/LCOMM.2019.2961890.

[51] Guo, Y. and Cheung, N.-M. Efficient and deep person re-identification using multi-level similarity. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2335–2344 (2018). doi:10.1109/CVPR.2018.00248.

[52] Hadsell, R., Chopra, S., and LeCun, Y. Dimensionality reduction by learning an invariant mapping. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 1735–1742 (2006). doi:10.1109/CVPR.2006.100.

[53] Halperin, D., Hu, W., Sheth, A., and Wetherall, D. Predictable 802.11 packet delivery from wireless channel measurements. *SIGCOMM Computer Communication Review*, **40** (2010), 159–170. doi:10.1145/1851275.1851203.

[54] Halperin, D., Hu, W., Sheth, A., and Wetherall, D. Tool release: Gathering 802.11 n traces with channel state information. *ACM SIGCOMM Computer Communication Review*, **41** (2011), 53. doi:10.1145/1925861.1925870.

[55] Hamida, E. B. and Chelius, G. Investigating the impact of human activity on the performance of wireless networks—an experimental approach. In *IEEE International Symposium on" A World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, pp. 1–8 (2010). doi:10.1109/WOWMOM.2010.5534913.

[56] Hoang, M. T., Yuen, B., Dong, X., Lu, T., Westendorp, R., and Reddy, K. Recurrent neural networks for accurate rssi indoor localization. *IEEE Internet of Things Journal*, **6** (2019), 10639. `doi:10.1109/JIOT.2019.2940368`.

[57] Hochreiter, S. and Schmidhuber, J. Long Short-Term Memory. *Neural Computation*, **9** (1997), 1735. `doi:10.1162/neco.1997.9.8.1735`.

[58] Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., and Chen, X. Vrstc: Occlusion-free video person re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7183–7192 (2019). `doi:10.1109/CVPR.2019.00735`.

[59] Huang, T., Yang, G., and Tang, G. A fast two-dimensional median filtering algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **27** (1979), 13. `doi:10.1109/TASSP.1979.1163188`.

[60] Huang, Y., Xu, J., Wu, Q., Zhong, Y., Zhang, P., and Zhang, Z. Beyond scalar neuron: Adopting vector-neuron capsules for long-term person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, **30** (2019), 3459. `doi:10.1109/TCSVT.2019.2948093`.

[61] Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, vol. 37, pp. 448–456. PMLR (2015).

[62] Isola, P., Zhu, J., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976 (2017). `doi:10.1109/CVPR.2017.632`.

[63] Jayasundara, V., Jayasekara, H., Samarasinghe, T., and Hemachandra, K. T. Device-free user authentication, activity classification and tracking using passive wi-fi sensing: A deep learning-based approach. *IEEE Sensors Journal*, **20** (2020), 9329. `doi:10.1109/JSEN.2020.2987386`.

[64] Jiang, H., Cai, C., Ma, X., Yang, Y., and Liu, J. Smart home based on wifi sensing: A survey. *IEEE Access*, **6** (2018), 13317. `doi:10.1109/ACCESS.2018.2812887`.

[65] Jiang, W., et al. Towards 3d human pose construction using wifi. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking (MobiCom '20)* (2020). `doi:10.1145/3372224.3380900`.

[66] Karanam, C. R. and Mostofi, Y. 3d through-wall imaging with unmanned aerial vehicles using wifi. In *ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pp. 131–142 (2017). `doi:10.1145/3055031.3055084`.

[67] Kato, S., Fukushima, T., Murakami, T., Abeysekera, H., Iwasaki, Y., Fujihashi, T., Watanabe, T., and Saruwatari, S. Csi2image: Image

reconstruction from channel state information using generative adversarial networks. *IEEE Access*, **9** (2021), 47154. `doi:10.1109/ACCESS.2021.3066158`.

[68] Kato, S., Fukushima, T., Murakami, T., Abeysekera, H., Iwasaki, Y., Fujihashi, T., Watanabe, T., and Saruwatari, S. Csi2image: Image reconstruction from channel state information using generative adversarial networks. *IEEE Access*, **9** (2021), 47154. `doi:10.1109/ACCESS.2021.3066158`.

[69] Kefayati, M. H., Pourahmadi, V., and Aghaeinia, H. Wi2vi: Generating video frames from wifi csi samples. *IEEE Sensors Journal*, **20** (2020), 11463. `doi:10.1109/JSEN.2020.2996078`.

[70] Kiefer, J. and Wolfowitz, J. Stochastic Estimation of the Maximum of a Regression Function. *The Annals of Mathematical Statistics*, **23** (1952), 462. `doi:10.1214/aoms/1177729392`.

[71] Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, pp. 1–15. Elsevier (2015).

[72] Leng, Q., Ye, M., and Tian, Q. A survey of open-world person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, **30** (2019), 1092. `doi:10.1109/TCSVT.2019.2898940`.

[73] Li, C., Cao, Z., and Liu, Y. Deep ai enabled ubiquitous wireless sensing: A survey. *ACM Computing Surveys*, **54** (2021), 1. `doi:10.1145/3436729`.

[74] Li, C., Liu, M., and Cao, Z. Wihf: Enable user identified gesture recognition with wifi. In *IEEE Conference on Computer Communications*, pp. 586–595 (2020). `doi:10.1109/INFOCOM41043.2020.9155539`.

[75] Li, H., Chen, Y., Tao, D., Yu, Z., and Qi, G. Attribute-aligned domain-invariant feature learning for unsupervised domain adaptation person re-identification. *IEEE Transactions on Information Forensics and Security*, **16** (2020), 1480. `doi:10.1109/TIFS.2020.3036800`.

[76] Li, H., He, X., Chen, X., Fang, Y., and Fang, Q. Wi-motion: A robust human activity recognition using wifi signals. *IEEE Access*, **7** (2019), 153287. `doi:10.1109/ACCESS.2019.2948102`.

[77] Li, J., Wang, J., Tian, Q., Gao, W., and Zhang, S. Global-local temporal representations for video person re-identification. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3958–3967 (2019). `doi:10.1109/ICCV.2019.00406`.

[78] Li, Q., Qu, H., Liu, Z., Zhou, N., Sun, W., Sigg, S., and Li, J. Af-dcgan: Amplitude feature deep convolutional gan for fingerprint construction in indoor localization systems. *IEEE Transactions on Emerging Topics in Computing*, **0** (2019), 1. `doi:10.1109/TETCI.2019.2948058`.

[79] Li, Q., Qu, H., Liu, Z., Zhou, N., Sun, W., Sigg, S., and Li, J. Af-dcgan: Amplitude feature deep convolutional gan for fingerprint construction in indoor localization systems. *IEEE Transactions on Emerging Topics in Computational Intelligence*, **5** (2021), 468. doi:10.1109/TETCI.2019.2948058.

[80] Li, W., Zhang, Z., and Liu, Z. Action recognition based on a bag of 3d points. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pp. 9–14 (2010). doi:10.1109/CVPRW.2010.5543273.

[81] Li, X., Li, S., Zhang, D., Xiong, J., Wang, Y., and Mei, H. Dynamic-music: Accurate device-free indoor localization. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*, p. 196–207 (2016). doi:10.1145/2971648.2971665.

[82] Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.-Y., and Kot, A. C. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **42** (2020), 2684. doi:10.1109/TPAMI.2019.2916873.

[83] Liu, S., Si, T., Hao, X., and Zhang, Z. Semantic constraint gan for person re-identification in camera sensor networks. *IEEE Access*, **7** (2019), 176257. doi:10.1109/ACCESS.2019.2958126.

[84] Liu, X., Tan, H., Tong, X., Cao, J., and Zhou, J. Feature preserving gan and multi-scale feature enhancement for domain adaption person re-identification. *Neurocomputing*, **364** (2019), 108 . doi:10.1016/j.neucom.2019.07.063.

[85] Ma, Y., Zhou, G., and Wang, S. Wifi sensing with channel state information: A survey. *ACM Computing Surveys*, **52** (2019), 1. doi:10.1145/3310194.

[86] Ma, Y., Zhou, G., and Wang, S. Wifi sensing with channel state information: A survey. *ACM Computing Surveys*, **52** (2019), 1. doi:10.1145/3310194.

[87] Maas, A., Hannun, A., and Ng, A. Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning (ICML)*, vol. 30, pp. 1–6. PMLR (2013).

[88] Mang, Y., Xiangyuan, L., Jiawei, L., and Pong C., Y. Hierarchical discriminative learning for visible thermal person re-identification. In *Conference on Artificial Intelligence, (AAAI)*, pp. 7501–7508 (2018).

[89] Marriott, R. T., Romdhani, S., and Chen, L. A 3d gan for improved large-pose facial recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13445–13455. IEEE (2021). doi:10.1109/cvpr46437.2021.01324.

[90] Mejjati, Y. A., Richardt, C., Tompkin, J., Cosker, D., and Kim, K. I. Unsupervised attention-guided image-to-image translation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, p. 3697–3707 (2018). doi:10.5555/3327144.3327286.

[91] Miao, J., Wu, Y., Liu, P., Ding, Y., and Yang, Y. Pose-guided feature alignment for occluded person re-identification. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 542–551 (2019). doi:10.1109/ICCV.2019.00063.

[92] Nagrani, A., Sun, C., Ross, D., Sukthankar, R., Schmid, C., and Zisserman, A. Speech2action: Cross-modal supervision for action recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10314–10323 (2020). doi:10.1109/CVPR42600.2020.01033.

[93] Nakamura, T., Bouazizi, M., Yamamoto, K., and Ohtsuki, T. Wi-fi-csi-based fall detection by spectrogram analysis with cnn. In *IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6 (2020). doi:10.1109/GLOBECOM42002.2020.9322323.

[94] Navaneet, K., Sarvadevabhatla, R. K., Shekhar, S., Babu, R. V., and Chakraborty, A. Operator-in-the-loop deep sequential multi-camera feature fusion for person re-identification. *IEEE Transactions on Information Forensics and Security*, **15** (2019), 2375. doi:10.1109/TIFS.2019.2957701.

[95] Nepovinnykh, E., Eerola, T., and Kalviainen, H. Siamese network based pelage pattern matching for ringed seal re-identification. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WAVC)*, pp. 25–34 (2020). doi:10.1109/WACVW50321.2020.9096935.

[96] Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., and Bajcsy, R. Berkeley mhad: A comprehensive multimodal human action database. In *IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 53–60 (2013). doi:10.1109/WACV.2013.6474999.

[97] Qian, K., Wu, C., Yang, Z., Liu, Y., and Jamieson, K. Widar: Decimeter-level passive tracking via velocity monitoring with commodity wi-fi. In *Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing* (2017). doi:10.1145/3084041.3084067.

[98] Qian, K., Wu, C., Yang, Z., Liu, Y., and Zhou, Z. Pads: Passive detection of moving targets with dynamic speed using phy layer information. In *IEEE 20th International Conference on Parallel and Distributed Systems (ICPADS)*, pp. 1–8 (2014). doi:10.1109/PADSW.2014.7097784.

[99] Qian, K., Wu, C., Yang, Z., Liu, Y., and Zhou, Z. Pads: Passive detection of moving targets with dynamic speed using phy layer information. In *IEEE International Conference on Parallel and Distributed Systems (ICPADS)*, pp. 1–8 (2014). doi:10.1109/PADSW.2014.7097784.

[100] Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations (ICLR*, pp. 1–16. Elsevier (2016).

[101] Rafael-Palou, X., Aubanell, A., Bonavita, I., Ceresa, M., Piella, G., Ribas, V., and Ballester, M. A. G. Re-identification and growth detection of pulmonary nodules without image registration using 3d siamese neural networks. *Medical Image Analysis*, **67** (2021), 101823. `doi:10.1016/j.media.2020.101823`.

[102] Rao, X., Li, Z., Yang, Y., and Wang, S. Dfphasefl: a robust device-free passive fingerprinting wireless localization system using csi phase information. *Neural Computing and Applications*, **32** (2020), 14909. `doi:10.1007/s00521-020-04847-1`.

[103] Regmi, K. and Borji, A. Cross-view image synthesis using geometry-guided conditional gans. *Computer Vision and Image Understanding*, **187** (2019), 102788. `doi:10.1016/j.cviu.2019.07.008`.

[104] Restuccia, F. IEEE 802.11bf: Toward ubiquitous wi-fi sensing. *arXiv Preprint*, **abs/2103.14918** (2021), 1. `doi:arXiv:2103.14918`.

[105] Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241 (2015). `doi:10.1007/978-3-319-24574-4_28`.

[106] Sangkloy, P., Lu, J., Fang, C., Yu, F., and Hays, J. Scribbler: Controlling deep image synthesis with sketch and color. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6836–6845 (2017). `doi:10.1109/CVPR.2017.723`.

[107] Schuldt, C., Laptev, I., and Caputo, B. Recognizing human actions: a local svm approach. In *International Conference on Pattern Recognition (ICPR)*, vol. 3, pp. 32–36 Vol.3 (2004). `doi:10.1109/ICPR.2004.1334462`.

[108] Sen, S., Radunovic, B., Choudhury, R. R., and Minka, T. You are facing the mona lisa: Spot localization using phy layer information. In *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, p. 183–196 (2012). `doi:10.1145/2307636.2307654`.

[109] Sharma, M. Novel adaptive channel state information feedback for multiuser mimo in wireless broadband communications. In *IEEE 14th International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*, pp. 1–2 (2013). `doi:10.1109/WoWMoM.2013.6583428`.

[110] Shen, X., Guo, L., Lu, Z., Wen, X., and He, Z. Wirim: Resolution improving mechanism for human sensing with commodity wi-fi. *IEEE Access*, **7** (2019), 168357. `doi:10.1109/ACCESS.2019.2954651`.

[111] Shen, X., Guo, L., Lu, Z., Wen, X., and He, Z. Wirim: Resolution improving mechanism for human sensing with commodity wi-fi. *IEEE Access*, **7** (2019), 168357. `doi:10.1109/ACCESS.2019.2954651`.

[112] Shi, S., Sigg, S., Chen, L., and Ji, Y. Accurate location tracking from csi-based passive device-free probabilistic fingerprinting. *IEEE Transactions on Vehicular Technology*, **67** (2018), 5217. `doi:10.1109/TVT.2018.2810307`.

[113] Shin, W., Bu, S.-J., and Cho, S.-B. 3d-convolutional neural network with generative adversarial network and autoencoder for robust anomaly detection in video surveillance. *International Journal of Neural Systems*, **30** (2020), 2050034. `doi:10.1142/S0129065720500343`.

[114] Shin, W., Bu, S.-J., and Cho, S.-B. 3d-convolutional neural network with generative adversarial network and autoencoder for robust anomaly detection in video surveillance. *International Journal of Neural Systems*, **30** (2020), 2050034. `doi:10.1142/S0129065720500343`.

[115] Simonyan, K. and ndrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, pp. 1–14 (2015).

[116] Song, C., Huang, Y., Ouyang, W., and Wang, L. Mask-guided contrastive attention model for person re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1179–1188 (2018). `doi:10.1109/CVPR.2018.00129`.

[117] Song, Q., Guo, S., Liu, X., and Yang, Y. Csi amplitude fingerprinting-based nb-iot indoor localization. *IEEE Internet of Things Journal*, **5** (2017), 1494. `doi:10.1109/JIOT.2017.2782479`.

[118] Song, S., Zhang, W., Liu, J., and Mei, T. Unsupervised person image generation with semantic parsing transformation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2352–2361 (2019). `doi:10.1109/CVPR.2019.00246`.

[119] Soro, B. and Lee, C. Joint time-frequency rssi features for convolutional neural network-based indoor fingerprinting localization. *IEEE Access*, **7** (2019), 104892. `doi:10.1109/ACCESS.2019.2932469`.

[120] Springenberg, J., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: The all convolutional net. In *International Conference on Learning Representations (ICLR)*, pp. 1–14. Elsevier (2015).

[121] Storrer, L., Yildirim, H. C., Crauwels, M., Copa, E. I. P., Pollin, S., Louveaux, J., De Doncker, P., and Horlin, F. Indoor tracking of multiple individuals with an 802.11ax wi-fi-based multi-antenna passive radar. *IEEE Sensors Journal*, **21** (2021), 20462. `doi:10.1109/JSEN.2021.3095675`.

[122] Stutzman, W. L. and Thiele, G. A. *Antenna Theory and Design*. John Wiley & Sons, 3rd edn. (2012). ISBN 9780470576649.

[123] Sun, X. and Zheng, L. Dissecting person re-identification from the viewpoint of viewpoint. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 608–617 (2019). `doi:10.1109/CVPR.2019.00070`.

[124] Sun, Y., Chen, Y., Wang, X., and Tang, X. Deep learning face representation by joint identification-verification. In *International Conference on Neural Information Processing System*, pp. 1988–1996 (2014). `doi:10.5555/2969033.2969049`.

[125] Sutskever, I., Martens, J., Dahl, G., and Hinton, G. On the importance of initialization and momentum in deep learning. In *Proceedings of Machine Learning Research (PMLR)*, pp. 1139–1147. PMLR (2013).

[126] Tagore, N. K., Singh, A., Manche, S., and Chattopadhyay, P. Person re-identification from appearance cues and deep siamese features. *Journal of Visual Communication and Image Representation*, **75** (2021), 103029. `doi:10.1016/j.jvcir.2021.103029`.

[127] Tong, X., Li, H., Tian, X., and Wang, X. Wi-fi localization enabling self-calibration. *IEEE/ACM Transactions on Networking*, **29** (2021), 904. `doi:10.1109/TNET.2021.3051998`.

[128] Tse, D. and Viswanath, P. *Fundamentals of Wireless Communication*. Cambridge University Press (2005). ISBN 9780521845274.

[129] Tse, D. and Viswanath, P. *Fundamentals of Wireless Communication*. Cambridge University Press (2005). ISBN 0521845270.

[130] Turaga, P. and Ivanov, Y. A. Diamond sentry: Integrating sensors and cameras for real-time monitoring of indoor spaces. *IEEE Sensors Journal*, **11** (2011), 593. `doi:10.1109/JSEN.2010.2050309`.

[131] Turaga, P. and Ivanov, Y. A. Diamond sentry: Integrating sensors and cameras for real-time monitoring of indoor spaces. *IEEE Sensors Journal*, **11** (2011), 593. `doi:10.1109/JSEN.2010.2050309`.

[132] Vlavianos, A., Law, L. K., Broustis, I., Krishnamurthy, S. V., and Faloutsos, M. Assessing link quality in ieee 802.11 wireless networks: Which is the right metric? In *IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 1–6 (2008). `doi:10.1109/PIMRC.2008.4699837`.

[133] Wang, F., Zhang, F., Wu, C., Wang, B., and Liu, K. R. Respiration tracking for people counting and recognition. *IEEE Internet of Things Journal*, **7** (2020), 5233. `doi:10.1109/JIOT.2020.2977254`.

[134] Wang, F., Zhou, S., Panev, S., Han, J., and Huang, D. Person-in-wifi: Fine-grained person perception using wifi. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5451–5460 (2019). `doi:10.1109/ICCV.2019.00555`.

[135] Wang, H., Zhang, D., Wang, Y., Ma, J., Wang, Y., and Li, S. Rt-fall: A real-time and contactless fall detection system with commodity wifi devices. *IEEE Transactions on Mobile Computing*, **16** (2016), 511. `doi:10.1109/TMC.2016.2557795`.

[136] Wang, J., Xiong, J., Jiang, H., Jamieson, K., Chen, X., Fang, D., and Wang, C. Low human-effort, device-free localization with fine-grained subcarrier information. *IEEE Transactions on Mobile Computing*, **17** (2018), 2550. `doi:10.1109/TMC.2018.2812746`.

[137] Wang, L., Chen, W., Yang, W., Bi, F., and Yu, F. R. A state-of-the-art review on image synthesis with generative adversarial networks. *IEEE Access*, **8** (2020), 63514. `doi:10.1109/ACCESS.2020.2982224`.

[138] Wang, W., Liu, A. X., Shahzad, M., Ling, K., and Lu, S. Device-free human activity recognition using commercial wifi devices. *IEEE Journal on Selected Areas in Communications*, **35** (2017), 1118. `doi:10.1109/JSAC.2017.2679658`.

[139] Wang, X., Gao, L., and Mao, S. Phasefi: Phase fingerprinting for indoor localization with a deep learning approach. In *IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6 (2015). `doi:10.1109/GLOCOM.2015.7417517`.

[140] Wang, X. and Yan, W. Q. Human gait recognition based on frame-by-frame gait energy images and convolutional long short-term memory. *International Journal of Neural Systems*, **30** (2020), 1950027. `doi:10.1142/S0129065719500278`.

[141] Wang, Y., Wu, K., and Ni, L. M. Wifall: Device-free fall detection by wireless networks. *IEEE Transactions on Mobile Computing*, **16** (2016), 581. `doi:10.1109/INFOCOM.2014.6847948`.

[142] Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, **13** (2004), 600. `doi:10.1109/TIP.2003.819861`.

[143] Wei, B., Hu, W., Yang, M., and Chou, C. T. From real to complex: Enhancing radio-based activity recognition using complex-valued csi. *ACM Transactions on Sensor Networks*, **15** (2019). `doi:10.1145/3338026`.

[144] Weinstein, S. B. The history of orthogonal frequency-division multiplexing [history of communications]. *IEEE Communications Magazine*, **47** (2009), 26. `doi:10.1109/MCOM.2009.5307460`.

[145] Wu, A., Zheng, W.-S., and Lai, J.-H. Robust depth-based person re-identification. *IEEE Transactions on Image Processing*, **26** (2017), 2588. `doi:10.1109/TIP.2017.2675201`.

[146] Wu, D., Zhang, D., Xu, C., Wang, H., and Li, X. Device-free wifi human sensing: From pattern-based to model-based approaches. *IEEE Communications Magazine*, **55** (2017), 91. `doi:10.1109/MCOM.2017.1700143`.

[147] Wu, D., Zhang, D., Xu, C., Wang, H., and Li, X. Device-free wifi human sensing: From pattern-based to model-based approaches. *IEEE Communications Magazine*, **55** (2017), 91. `doi:10.1109/MCOM.2017.1700143`.

[148] Wu, J., Zhang, C., Xue, T., Freeman, W. T., and Tenenbaum, J. B. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, p. 82–90 (2016). doi:10.5555/3157096.3157106.

[149] Wu, K., Xiao, J., Yi, Y., Chen, D., Luo, X., and Ni, L. M. Csi-based indoor localization. *IEEE Transactions on Parallel and Distributed Systems*, **24** (2013), 1300. doi:10.1109/TPDS.2012.214.

[150] Wu, L., Wang, Y., Gao, J., and Li, X. Where-and-when to look: Deep siamese attention networks for video-based person re-identification. *IEEE Transactions on Multimedia*, **21** (2018), 1412. doi:10.1109/TMM.2018.2877886.

[151] Wu, X., Chu, Z., Yang, P., Xiang, C., Zheng, X., and Huang, W. Tw-see: Human activity recognition through the wall with commodity wi-fi devices. *IEEE Transactions on Vehicular Technology*, **68** (2018), 306. doi:10.1109/TVT.2018.2878754.

[152] Wu, X., Chu, Z., Yang, P., Xiang, C., Zheng, X., and Huang, W. Tw-see: Human activity recognition through the wall with commodity wi-fi devices. *IEEE Transactions on Vehicular Technology*, **68** (2019), 306. doi:10.1109/TVT.2018.2878754.

[153] Wu, X., Xu, K., and Hall, P. A survey of image synthesis and editing with generative adversarial networks. *Tsinghua Science and Technology*, **22** (2017), 660. doi:10.23919/TST.2017.8195348.

[154] Xiao, F., Chen, J., Xie, X., Gui, L., Sun, L., and Wang, R. Seare: A system for exercise activity recognition and quality evaluation based on green sensing. *IEEE Transactions on Emerging Topics in Computing*, **8** (2020), 752. doi:10.1109/TETC.2018.2790080.

[155] Xiao, F., Chen, J., Xie, X. H., Gui, L., Sun, J. L., and none Ruchuan, W. Seare: A system for exercise activity recognition and quality evaluation based on green sensing. *IEEE Transactions on Emerging Topics in Computational Intelligence*, **8** (2018), 752. doi:10.1109/TETC.2018.2790080.

[156] Xiao, J., Wu, K., Yi, Y., Wang, L., and Ni, L. M. Pilot: Passive device-free indoor localization using channel state information. In *IEEE Int. Conf. Distr. Comput. Syst. (ICDCS)*, pp. 236–245 (2013). doi:10.1109/ICDCS.2013.49.

[157] Xie, Y., Xiong, J., Li, M., and Jamieson, K. Md-track: Leveraging multi-dimensionality for passive indoor wi-fi tracking. In *The 25th Annual International Conference on Mobile Computing and Networking (MobiCom '19*, pp. 1–16 (2019). doi:10.1145/3300061.3300133.

[158] Xu, Q., Chen, Y., Wang, B., and Liu, K. J. R. Radio biometrics: Human recognition through a wall. *IEEE Transactions on Information Forensics and Security*, **12** (2017), 1141. `doi:10.1109/TIFS.2016.2647224`.

[159] Xu, Z., Wang, R., Yue, X., Liu, T., Chen, C., and Fang, S.-H. Faceme: Face-to-machine proximity estimation based on rssi difference for mobile industrial human–machine interaction. *IEEE Transactions on Industrial Informatics*, **14** (2018), 3547. `doi:10.1109/TII.2018.2829847`.

[160] Xue, W., Li, Q., Hua, X., Yu, K., Qiu, W., and Zhou, B. A new algorithm for indoor rssi radio map reconstruction. *IEEE Access*, **6** (2018), 76118. `doi:10.1109/ACCESS.2018.2882379`.

[161] Xue, W., et al. Eight-diagram based access point selection algorithm for indoor localization. *IEEE Transactions on Vehicular Technology*, **69** (2020), 13196. `doi:10.1109/TVT.2020.3021090`.

[162] Yan, M., Jiang, X., and Yuan, J. 3d convolutional generative adversarial networks for detecting temporal irregularities in videos. In *International Conference on Pattern Recognition (ICPR)*, pp. 2522–2527. IEEE (2018). `doi:10.1109/icpr.2018.8546039`.

[163] Yang, J., Zou, H., Zhou, Y., and Xie, L. Learning gestures from wifi: A siamese recurrent convolutional architecture. *IEEE Internet of Things Journal*, **6** (2019), 10763. `doi:10.1109/JIOT.2019.2941527`.

[164] Yang, J., Zou, H., Zhou, Y., and Xie, L. Learning gestures from wifi: A siamese recurrent convolutional architecture. *IEEE Internet of Things Journal*, **6** (2019), 10763. `doi:10.1109/JIOT.2019.2941527`.

[165] Yang, Z., Zhou, Z., and Liu, Y. From rssi to csi: Indoor localization via channel response. *ACM Computing Surveys*, **46** (2013), 1. `doi:10.1145/2543581.2543592`.

[166] Yang, Z., Zhou, Z., and Liu, Y. From rssi to csi: Indoor localization via channel response. *ACM Computing Surveys*, **46** (2013), 1. `doi:10.1145/2543581.2543592`.

[167] Yao, R., Gao, C., Xia, S., Zhao, J., Zhou, Y., and Hu, F. Gan-based person search via deep complementary classifier with center-constrained triplet loss. *Pattern Recognition*, **104** (2020), 107350. `doi:10.1016/j.patcog.2020.107350`.

[168] Ye, M., Lan, X., Wang, Z., and Yuen, P. C. Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE Transactions on Information Forensics and Security*, **15** (2019), 407. `doi:10.1109/TIFS.2019.2921454`.

[169] Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., and Hoi, S. C. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2021), 1. `doi:10.1109/TPAMI.2021.3054775`.

[170] Ye, M., Shen, J., and Shao, L. Visible-infrared person re-identification via homogeneous augmented tri-modal learning. *IEEE Transactions on Information Forensics and Security*, **16** (2020), 728. `doi:10.1109/TIFS.2020.3001665`.

[171] Ye, M., Shen, J., Zhang, X., Yuen, P. C., and Chang, S.-F. Augmentation invariant and instance spreading feature for softmax embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2020), 1. `doi:10.1109/TPAMI.2020.3013379`.

[172] Ye, M. and Yuen, P. C. Purifynet: A robust person re-identification model with noisy labels. *IEEE Transactions on Information Forensics and Security*, **15** (2020), 2655. `doi:10.1109/TIFS.2020.2970590`.

[173] Yu, B., Wang, Y., Niu, K., Zeng, Y., Gu, T., Wang, L., Guan, C., and Zhang, D. Wifi-sleep: Sleep stage monitoring using commodity wi-fi devices. *IEEE Internet of Things Journal*, **8** (2021), 13900. `doi:10.1109/JIOT.2021.3068798`.

[174] Yun, Z. and Iskander, M. F. Ray tracing for radio propagation modeling: Principles and applications. *IEEE Access*, **3** (2015), 1089. `doi:10.1109/ACCESS.2015.2453991`.

[175] Zhang, C., Zhu, L., Zhang, S., and Yu, W. Pac-gan: An effective pose augmentation scheme for unsupervised cross-view person re-identification. *Neurocomputing*, **387** (2020), 22–39. `doi:10.1016/j.neucom.2019.12.094`.

[176] Zhang, C., Zhu, L., Zhang, S., and Yu, W. Pac-gan: An effective pose augmentation scheme for unsupervised cross-view person re-identification. *Neurocomputing*, **387** (2020), 22 . `doi:10.1016/j.neucom.2019.12.094`.

[177] Zhang, F., Chen, C., Wang, B., and Liu, K. J. R. Wispeed: A statistical electromagnetic approach for device-free indoor speed estimation. *IEEE Internet of Things Journal*, **5** (2018), 2163. `doi:10.1109/JIOT.2018.2826227`.

[178] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5908–5916 (2017). `doi:10.1109/ICCV.2017.629`.

[179] Zhang, J., Tang, Z., Li, M., Fang, D., Nurmi, P., and Wang, Z. Crosssense: Towards cross-site and large-scale wifi sensing. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking (MobiCom '18)*, p. 305–320 (2018). `doi:10.1145/3241539.3241570`.

[180] Zhang, L., Zhang, L., Mou, X., and Zhang, D. Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, **20** (2011), 2378. `doi:10.1109/TIP.2011.2109730`.

[181] Zhang, S., Zhang, L., Wang, W., and Wu, X. Asnet: Asymmetrical network for learning rich features in person re-identification. *IEEE Signal Processing Letters*, **27** (2020), 850. `doi:10.1109/LSP.2020.2994815`.

[182] Zhang, Y., Zheng, Y., Qian, K., Zhang, G., Liu, Y., Wu, C., and Yang, Z. Widar3.0: Zero-effort cross-domain gesture recognition with wi-fi. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2021), 1. `doi:10.1109/TPAMI.2021.3105387`.

[183] Zhang, Z. and Peng, H. Deeper and wider siamese networks for real-time visual tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4586–4595 (2019). `doi:10.1109/CVPR.2019.00472`.

[184] Zhao, M., Li, T., Alsheikh, M. A., Tian, Y., Zhao, H., Torralba, A., and Katabi, D. Through-wall human pose estimation using radio signals. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7356–7365 (2018). `doi:10.1109/CVPR.2018.00768`.

[185] Zhao, M., Tian, Y., Zhao, H., Alsheikh, M. A., Li, T., Hristov, R., Kabelac, Z., Katabi, D., and Torralba, A. Rf-based 3d skeletons. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*, p. 267–281 (2018). `doi:10.1145/3230543.3230579`.

[186] Zheng, C., Cham, T.-J., and Cai, J. Pluralistic image completion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1438–1447 (2019). `doi:10.1109/CVPR.2019.00153`.

[187] Zheng, M., Karanam, S., Wu, Z., and Radke, R. J. Re-identification with consistent attentive siamese networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5735–5744 (2019). `doi:10.1109/CVPR.2019.00588`.

[188] Zheng, X., Wang, J., Shangguan, L., Zhou, Z., and Liu, Y. Design and implementation of a csi-based ubiquitous smoking detection system. *IEEE/ACM Transactions on Networking*, **25** (2017), 3781. `doi:10.1109/TNET.2017.2752367`.

[189] Zheng, Z., Zheng, L., and Yang, Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 3774–3782 (2017). `doi:10.1109/ICCV.2017.405`.

[190] Zhou, S., Ke, M., and Luo, P. Multi-camera transfer gan for person re-identification. *Journal of Visual Communication and Image Representation*, **59** (2019), 393 . `doi:10.1016/j.jvcir.2019.01.029`.

[191] Zhou, S., Wang, F., Huang, Z., and Wang, J. Discriminative feature learning with consistent attention regularization for person re-identification. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8040–8049 (2019). `doi:10.1109/ICCV.2019.00813`.

[192] Zhou, S., Wang, J., Meng, D., Xin, X., Li, Y., Gong, Y., and Zheng, N. Deep self-paced learning for person re-identification. *Pattern Recognition*, **76** (2018), 739. doi:10.1016/j.patcog.2017.10.005.

[193] Zhou, X., Huang, S., Li, B., Li, Y., Li, J., and Zhang, Z. Text guided person image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3658–3667 (2019). doi:10.1109/CVPR.2019.00378.

[194] Zhou, Z., Wu, C., Yang, Z., and Liu, Y. Sensorless sensing with wifi. *Tsinghua Science and Technology*, **20** (2015), 1. doi:10.1109/TST.2015.7040509.

[195] Zhu, B. and Ngo, C.-W. Cookgan: Causality based text-to-image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5518–5526 (2020). doi:10.1109/CVPR42600.2020.00556.

[196] Zhu, J., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251 (2017). doi:10.1109/ICCV.2017.244.