IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 September 2022, accepted 29 September 2022, date of publication 14 October 2022, date of current version 24 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3214972

RESEARCH ARTICLE

RF-Inpainter: Multimodal Image Inpainting Based on Vision and Radio Signals

CHENG CHEN[1], TAKAYUKI NISHIO[1], (Senior Member, IEEE), MEHDI BENNIS[2], (Fellow, IEEE), AND JIHONG PARK[3], (Senior Member, IEEE)

[1]School of Engineering, Tokyo Institute of Technology, Ookayama, Meguro-ku, Tokyo 152-8550, Japan
[2]Centre of Wireless Communications, University of Oulu, 90014 Oulu, Finland
[3]School of Info Technology, Deakin University, Geelong, VIC 3220, Australia

Corresponding author: Takayuki Nishio (nishio@ict.e.titech.ac.jp)

This work was supported in part by JSPS KAKENHI under Grant JP22H03575.

ABSTRACT This study demonstrates the feasibility of image inpainting using both visual information and radio frequency (RF) signals. Recent developments in imaging and vision-based technologies using RF signals have revealed the potential of leveraging multimodal information to enhance image inpainting performance. In this context, we propose RF-Inpainter—a novel inpainting method that integrates visual and wireless information by fusing defective RGB images with received signal strength indicator (RSSI) using a deep auto-encoder model. The inpainting performance of RF-Inpainter is evaluated using experimentally obtained images and RSSI datasets in an indoor environment. Image-only inpainting and RSSI-only inpainting models are used as baselines to illustrate the superiority of RF-Inpainter over inpainting methods based on a single modality. The results establish that RF-Inpainter generates satisfactory inpainted images in most experimental scenarios, achieving a maximum improvement of 36.4% and 14.6% in terms of mean peak signal-to-noise ratio (PSNR) and mean structural similarity index (SSIM), respectively.

INDEX TERMS Image inpainting, multi-modal, WiFi sensing, deep learning, RSSI fingerprint.

I. INTRODUCTION

The goal of image inpainting is to repair missing or damaged areas in defective images to match the original content as closely as possible. Image inpainting techniques have a wide range of applications in computer vision (e.g., [1], [2], and [3]) and public safety maintenance scenarios which involve the deployment of surveillance devices in public places to protect monitored areas or investigate crimes. Almost all existing image inpainting methods are based on a common principle—they use pixels from uncorrupted image regions to fill the gaps, similar to physical inpainting methods [4]. However, although these techniques are effective for recovering small-sized missing areas, they are not effective corresponding to very large missing areas, owing to the small amount of visual information remaining in the severely defective images. In this case, other information must be sued for image inpainting.

RF signal-based imaging has been suggested as a novel solution to this problem. Radar sensing is among the most traditional methods used to obtain images from RF signals—radar antennas are used to determine the distance between the antenna and the reflecting object, the amplitude of the echo, and its phase by transmitting thousands of pulses of microwave radiation and measuring the characteristics of the associated echoes [5]. By processing and combining these measurements together, images can be acquired. Over the last decade, various imaging methods based on wireless communication signals, such as WiFi, have been proposed [6], [7], [8], [9]. The successful acquisition of images from wireless information indicates the feasibility of image inpainting using RF information. However, the most appropriate types of RF information for image inpainting remain to be determined. Some studies have used commercial radar systems to obtain high-resolution images of objects, but constructing such

The associate editor coordinating the review of this manuscript and approving it for publication was Adam Czajka.

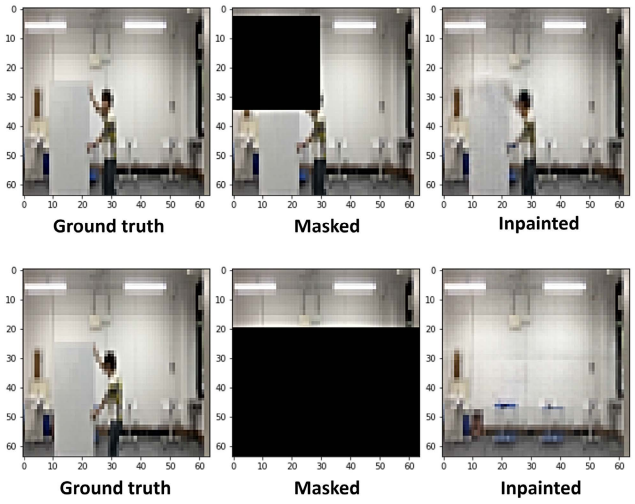

FIGURE 1. The existing inpainting techniques are not effective corresponding to very large missing regions.

complex systems is often expensive [10], [11]. In contrast, the use of ready-made wireless information enables accessible imaging at the cost of clarity of the imaging procedure. Recently, several studies have demonstrated that the channel state information (CSI) or received signal strength indicator (RSSI) of RF signals can be used as reliable sources of visual information (e.g., [12], [13], [14], and [15]). These features contain visual clues about defective regions in images, which provide prior information for image inpainting and improve its accuracy.

In this study, we propose RF-Inpainter—a novel inpainting method that integrates visual and wireless information. The fundamental operational idea is to use multimodal deep learning to fuse defective RGB images with RSSI of WiFi signals to produce inpainted images. We utilize WiFi signals because of their ubiquity. RSSI fingerprints are employed as wireless information because they can be easily measured using off-the-shelf electronic devices, such as laptops and smartphones. Although CSI contains greater propagation information, it cannot be obtained using commercial devices. The proposed method can also be employed based on CSI, yielding better inpainting results—we relegate this investigation to future works.

The primary contributions of this study are summarized as follows:

- A novel method called RF-Inpainter is proposed to perform high-resolution image inpainting based on heterogeneous modalities, namely vision and RF signals. Unlike traditional methods that rely solely on visual information for image inpainting or RF information for imaging, RF-Inpainter leverages both undamaged pixels surrounding a target missing region and RSSI to obtain inpainted images.
- The feasibility of image inpainting using RF-Inpainter is demonstrated using two experimentally obtained multimodal datasets. Moreover, the superiority of RF-Inpainter over inpainting methods leveraging a single modality is demonstrated by comparing its performance to those of Image-only and RSSI-only inpainting models.
- Experimental results reveal the relationship between the length of RSSI vectors and the image inpainting performance of RF-Inpainter— the quality of inpainted image is observed to be a function of RSSI vector length.

The remainder of the paper is organized as follows: Section II elaborates on related works, followed by Section III, where we introduce the image inpainting mechanism based on RF-Inpainter. Section IV evaluates the image inpainting results under different occlusion scenarios using Camera 1 and Camera 2 image datasets. Finally, the conclusions are presented in Section V.

II. RELATED WORKS

A. INPAINTING LARGE MISSING REGIONS BASED ON VISUAL INFORMATION

In computer vision, various image inpainting methods have been proposed to fill in defective regions in images to match ground-truth images as closely as possible. Several comprehensive reviews of image inpainting methods have been reported over the past decade [4], [16]. These methods can be classified into two primary categories—deep learning-based methods that are currently widely used, and traditional non-deep learning-based methods.

Traditional methods are suitable for filling in small or regular-sized missing regions in images. However, the inpainting task becomes more difficult when larger regions need to be filled in, irrespective of their position in the image. Traditional methods are not capable of filling in large complex regions satisfactorily—this has motivated the development of deep learning-based methods.

In recent years, deep learning-based methods have performed exceptionally in filling in large missing regions in image-inpainting tasks. For instance, Ma et al. proposed a generic inpainting framework capable of inpainting incomplete images containing both contiguous and discontiguous large missing areas based on generative adversarial networks (GANs) [17]. However, GAN is more suited to capture data distributions rather than image content or semantics. As a result, it may produce images that are significantly different from ground truth images [18]. To address this problem, Jia Q et al. presented a weighted face similarity (WFS)-Net-based face inpainting framework to improve inpainting performance [18]. First, they constructed a WFS set based on SSIM to gather a great amount of information for filling in missing regions. Subsequently, they designed a WFS-Net to inpaint damaged face images by exploring the relationship between missing regions and reference information including the remaining image parts and their WFSs. Li et al. proposed a recurrent learning-based approach to solve this problem [19].

They first mapped the RGB image to be repaired into a convolutional feature space. Next, they employed a recurrent feature reasoning (RFR) network to infer the boundaries of the missing parts in the feature maps repeatedly. Eventually, the feature maps were restored, combined, and transformed back into an RGB image.

Despite significant advancements in deep learning-based methods, filling in large missing image regions remains a challenging task. Although state-of-the-art methods have achieved satisfactory inpainting results, they can only restore semantically reasonable and visually realistic images. These methods infer missing content solely based on an image's residual information, inducing a difference between reconstructed image and ground truth image. This shortcoming is unacceptable for practical applications, such as criminal investigations and traffic cameras. In addition, these inpainting methods often leverage standard convolutional structures, which may lead to problems such as color discrepancies and blurring in reconstructed images [16].

B. IMAGING USING RF SIGNALS

Over the last decade, various techniques have been proposed for non-line-of-sight imaging using RF signals. These techniques overcome the drawbacks of traditional optical imaging (e.g., vulnerability to illumination conditions and occlusions). In [7], Vakalis et al. broadly classified all existing RF-based imaging systems into three categories—mechanical and electronic scanning imagers [20], holographic imaging systems [21], and staring-type imagers [22], [23]. However, a common drawback of these systems is that they cannot achieve cost-effective and real-time imaging. Therefore, Vakalis et al. developed a new microwave computational imaging system that requires a low receiver gain and does not rely on mechanical or electrical beam scanning [7]. Nevertheless, the construction of such an imaging system is time-consuming, and the imaging performance requires further improvements.

More recently, several studies have proposed RF imaging methods that can extract visual information from easy-to-acquire wireless signal features (e.g., RSSI or CSI) obtained by WiFi devices, which greatly simplifies the hardware structure of imaging system. For example, Kato et al. proposed a GAN-based technique called CSI2Image to reconstruct images using CSI. They also developed two applications—material sensing, and device-free user localization—to demonstrate the versatility of CSI2Image [12]. In [13], Dubey et al. developed an extended Rytov phaseless imaging (xRPI) technique that images object shapes and refractive indices in an indoor environment using RSSI of WiFi signals. These advanced imaging techniques have motivated novel ideas to improve image inpainting performance. By fusing defective RGB images with wireless features, such as RSSI, inpainted images in high resolution can be acquired under various situations.

C. FUSING WIRELESS SIGNALS WITH RGB IMAGES

The fusion of wireless signals and RGB images has driven technological advances in two tasks—vision to communication (V2C) and communication to vision (C2V) [24]. In V2C, several studies have demonstrated that the robustness of wireless communication systems can be significantly enhanced with the aid of visual information, such as RGB depth (RGB-D) images and light detection. In [25], we considered the scenario of a communication link randomly blocked by obstacles in an indoor environment, and we achieved more accurate prediction of millimeter-wave wireless channel dynamics, such as future received power and channel blockage, by transmitting RGB-D images into a deep neural network. Another example is a vision-assisted proactive handover framework [26]. With the help of time-continuous camera images, this study generated a better handover strategy based on a prior perception of obstacles by utilizing deep reinforcement learning to predict future transitions between line-of-sight and non-line-of-sight communication.

Recent studies on C2V have primarily focused on the integration of wireless and visual information to implement CV tasks, such as high-precision indoor localization and trajectory prediction. For example, Jiao et al. proposed a smartphone-based indoor positioning algorithm, in which wireless signals and RGB images are deeply fused to improve indoor human localization performance [27]. Zhu et al. proposed a system for identification and target tracking by combining wireless signals and computer vision [28]. The system achieves reliable trajectory matching and re-identification that can provide identity information of visual trajectories. The system also helps mitigate the effects of occlusion and illumination conditions on localization.

In the context of image inpainting, our previous work suggested carrying out research in RF-assisted image inpainting and demonstrated the possibility of performance improvement by fusing RF signals with RGB images in a preliminary experiment using mmWave communications, which is strongly attenuated by human blockage [24]. Building on the previous preliminary study, we propose an image inpainting method by integrating visual and wireless information in this study and demonstrate the feasibility of RF-assisted image inpainting in the lower frequency band (i.e., 5 GHz).

III. RF-INPAINTER: MULTIMODAL IMAGE INPAINTING

A. SYSTEM MODEL

We consider a scenario comprising multiple surveillance devices deployed in public places, such as corridors, security checkpoints, and bank counters, to monitor target areas. These sites are usually covered by wireless communication networks, such as WiFi systems—thus, we leverage the signals from multiple WiFi access points (APs) that are installed in the environment beforehand. When surveillance is active, the line of sight between the target area and the camera may be continuously obstructed by moving obstacles (e.g., pedestrians). This non-line-of-sight artifact results in partial

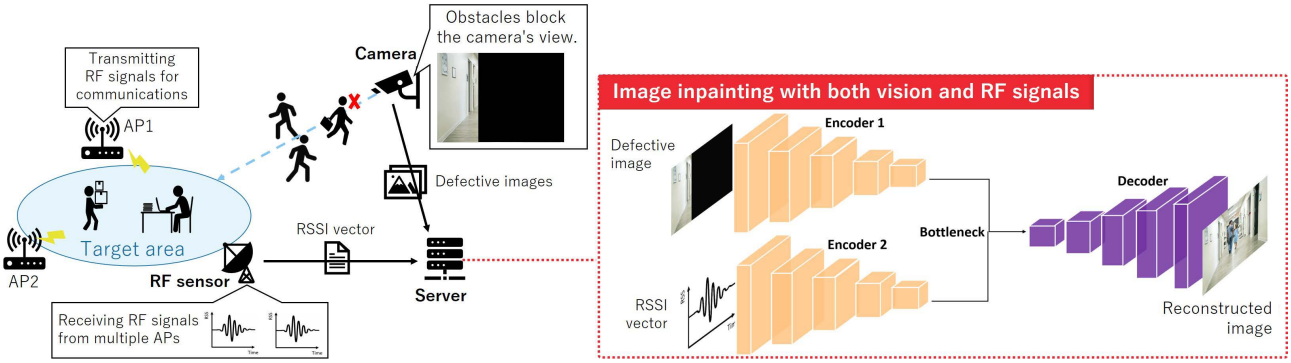

FIGURE 2. The system model of RF-Inpainter.

or complete absence of valuable information from captured images, which poses a severe problem in practical applications. In this case, we employ RF-Inpainter to perform inpainting tasks on captured defective images using WiFi signals already existing in the environment.

The system model of RF-Inpainter is depicted in Fig. 2. The system consists of APs that transmit WiFi signals, RGB cameras, RF sensors, and a server with an image-inpainting module based on heterogeneous modalities. RGB cameras capture real-time images of the target area. Simultaneously, RF sensors continuously measure the RSSI fluctuations of WiFi signals emitted from the APs. The images and RSSI data obtained via the RGB cameras and RF sensors, respectively, are uploaded to the server. The server detects images with missing regions and transmits them into the image inpainting module with their corresponding RSSI sequences to obtain distortion-free images.

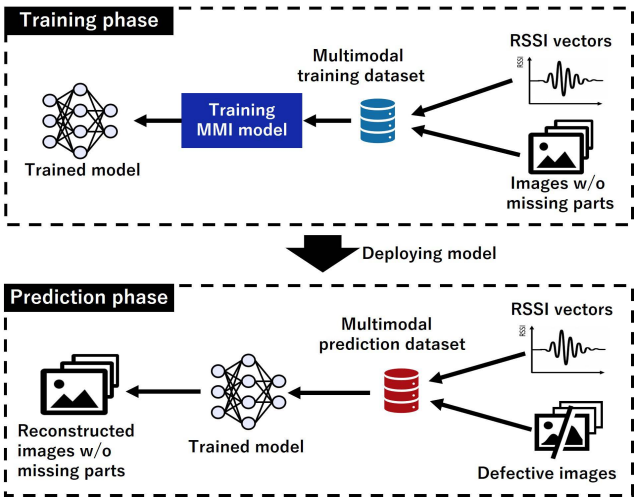

FIGURE 3. The inpainting procedure of RF-Inpainter.

B. RF-INPAINTER OVERVIEW

RF-Inpainter employs deep learning to enable image inpainting based on vision and radio signals. The deep learning model is called multimodal inpainting (MMI) model, which is further elaborated upon in Section II-E. The MMI model accepts a defective image and its corresponding RSSI values as inputs simultaneously and outputs a complete image by restoring the missing content.

The operations of the proposed RF-Inpainter can be roughly divided into model training and prediction (i.e., inpainting) phases. Fig. 3 illustrates the procedures of the training and prediction. First, model training is conducted on successfully obtained images. RF-Inpainter automatically constructs training data for the MMI model based on the observed RSSI and complete images and updates the MMI model, as detailed in Section II-D. Following model training, prediction is performed for observed images containing missing regions. Specifically, the trained MMI model takes a defective image and the corresponding RSSI values recorded in the same time window and generates a complete image without any missing regions.

RF-Inpainter performs image inpainting using RSSI of WiFi signals. WiFi is the most appropriate candidate for RF signals because of its ubiquity and low cost. In public places, such as offices, airports, and shopping malls, a large number of WiFi APs are usually already deployed, which can effectively improve the accuracy of image inpainting. Currently, several existing WiFi-based imaging techniques employ fine-grained features, such as CSI, for imaging by identifying the multi-path characteristics of radio channels [12], [14], [15]. Nevertheless, CSI can only be obtained using a few commercial devices, and its measurement requires a deep investigation of the PHY layer [29]. In contrast, RSSI not only characterizes the attenuation of radio signals during propagation but can also be measured using most off-the-shelf wireless devices—thus, the measurement of RSSI is faster and easier.

The key to inpainting with RF-Inpainter is imaging using RSSI of WiFi signals. The rationale for why RSSI can be used for imaging is that the strength of RSSI is impacted by the displacements and movements of the transmitter, receiver, and surrounding objects like humans. RSSI attenuates when

the wireless communication path is blocked by pedestrians or other obstacles. Based on the shape and location of different obstacles, the degree of attenuation of RSSI varies, which makes RSSI captures the wireless characteristics of the environment around the communication path. These characteristics, assisted by mathematical modeling or deep-learning algorithms, can be used for sensing applications like imaging. Since such attenuation of RSSI can only be described by multiple values, we need to prepare a temporally continuous sequence of RSSI (i.e., RSSI vectors) for each image to be restored. In addition, the acquisition moment of each image should be as close as possible to the acquisition moment of its corresponding RSSI vector.

Moreover, inspired by existing RSSI-based sensing techniques, we utilize a multi-AP network to improve imaging accuracy by fusing RSSI values obtained from different APs. The use of multiple APs expands signal coverage by including signals on different propagation paths, thereby avoiding the existence of blind spots and completely sensing the environment. Additionally, environmental factors (e.g., ubiquitous noise, changes in temperature and humidity, and variations in illumination intensity at different times of the day) might affect the accuracy of RSSI measurements, which can also be addressed by fusing RSSI of WiFi signals obtained from APs deployed at different locations.

C. DATA PREPROCESSING AND THE GENERATION OF MULTIMODAL DATASETS

As mentioned in Section II-A, the process of inpainting using RF-Inpainter can be broadly divided into two phases—model training and real-time prediction. During each phase, collected defective images and RSSI vectors are preprocessed to generate a multimodal dataset. As the data preprocessing methodologies in both phases are roughly identical, we focus on the generation of a multimodal training dataset and Fig. 4 summarizes this process. The steps involved are described below.

1) DATA ACQUISITION AND DOWNSAMPLING

Initially, the defective images and corresponding RSSI data are collected to train and validate the MMI model. We first sample clear and complete images and the corresponding RSSI values obtained from around the target area, and then upload the sampled data to the server. After that, various missing regions are artificially generated on the images to simulate that the camera's field of view is blocked by obstacles. During data acquisition, RSSI is sampled at a rate equal to or higher than that of RGB images to ensure that each image possesses sufficiently many corresponding RSSI values to constitute a vector. And we must ensure that the RSSI acquisition period and the image acquisition period roughly overlap.

Subsequently, each image is numbered in the order of its acquisition time to facilitate the sorting and deleting of images. The server then performs image downsampling to reduce the computational effort. To illustrate the downsampling process, let $t \in \mathbb{Z}$ denote a time index and $\boldsymbol{i}_t$ be the RGB image captured by a camera at time t. Suppose that the size of the RGB image is $H \times W$, where H and W denote the height and width of the image, respectively. After downsampling, the image size is compressed from $H \times W$ to $h \times w$ ($h \leq H, w \leq W$). We use $\boldsymbol{i'}_t$ to denote the downsampled $\boldsymbol{i}_t$. If s images are acquired between time 0 and time t in aggregate, the tensor $\boldsymbol{x}_t = [\boldsymbol{i'}_{t-s+1}, \boldsymbol{i'}_{t-s+2}, \ldots, \boldsymbol{i'}_t]$ can be employed to represent all downsampled images.

2) LABELING IMAGES WITH RSSI VECTORS

In this step, each image is labeled with multiple sequential RSSI values (i.e., an RSSI vector). Labeling enables the MMI model to learn the mapping from wireless information to vision.

Suppose $\boldsymbol{R}_t$ is the RSSI vector to be allocated as a label for the image $\boldsymbol{i'}_t$. To obtain $\boldsymbol{R}_t$, L ($L = 2l$) time-sequential RSSI values whose acquisition times are centered on t are required. In general, the value of L should not be too small nor too large—if L is too small, the RSSI vector does not contain sufficient spatial information about the environment; whereas if L is too large, it increases the computational complexity and introduces interference. In Section IV-F, we elaborate on the selection of the optimal value for L.

The RSSI value measured using the RF sensor at time t is denoted by y_t. Assuming that the sampling rate of the RSSI values is F, the time required to measure the RSSI value is $k = 1/F$ s. In this case, the RSSI vector, $\boldsymbol{R}_t$, used to label the image $\boldsymbol{i'}_t$, can be represented as $\boldsymbol{R}_t = [y_{t-lk}, y_{t-l(k-1)}, \ldots, y_t, \ldots, y_{t+l(k-2)}, y_{t+l(k-1)}]$, where $y_{t-lk}, \ldots y_{t-1}$ and $y_{t+1}, \ldots, y_{t+l(k-1)}$ denote past and future RSSI values based on y_t, respectively. Let T denote the index set of t corresponding to the captured samples. Then, the produced RSSI vector-image multimodal dataset can be represented as $D = \{\boldsymbol{i'}_t, \boldsymbol{R}_t | t \in T\}$.

When multiple RF sensors are deployed, the image, $\boldsymbol{i'}_t$, corresponds to multiple RSSI vectors. Assuming that there are n RF sensors, the RSSI values measured via each RF sensor at time t can be expressed as $y_t^1, y_t^2, \ldots, y_t^n$. According to the labeling method described above, we obtain $\boldsymbol{R}_t^1, \boldsymbol{R}_t^2, \ldots, \boldsymbol{R}_t^n$ as the corresponding RSSI vectors of $\boldsymbol{i'}_t$ from each RF sensor. Thus, the eventually generated RSSI vector–image multimodal dataset is expressed as $D^n = \{\boldsymbol{i'}_t, \boldsymbol{R}_t^1, \boldsymbol{R}_t^2, \ldots, \boldsymbol{R}_t^n | t \in T\}$.

3) MASKING RAW IMAGES TO GENERATE TRAINING DATA

Subsequently, we simulate the camera views being blocked by obstacles through masking the captured raw images to produce defective images for model training.

First, we employ two classical image-masking methods—horizontal and vertical masking—to remove all information from a fixed region of the image consistently and steadily. In addition, for more common situations, we configure random- occlusion scenarios by arbitrarily varying the size and position of the missing region in an image to imitate realistic scenarios in which various objects may block the cameras.

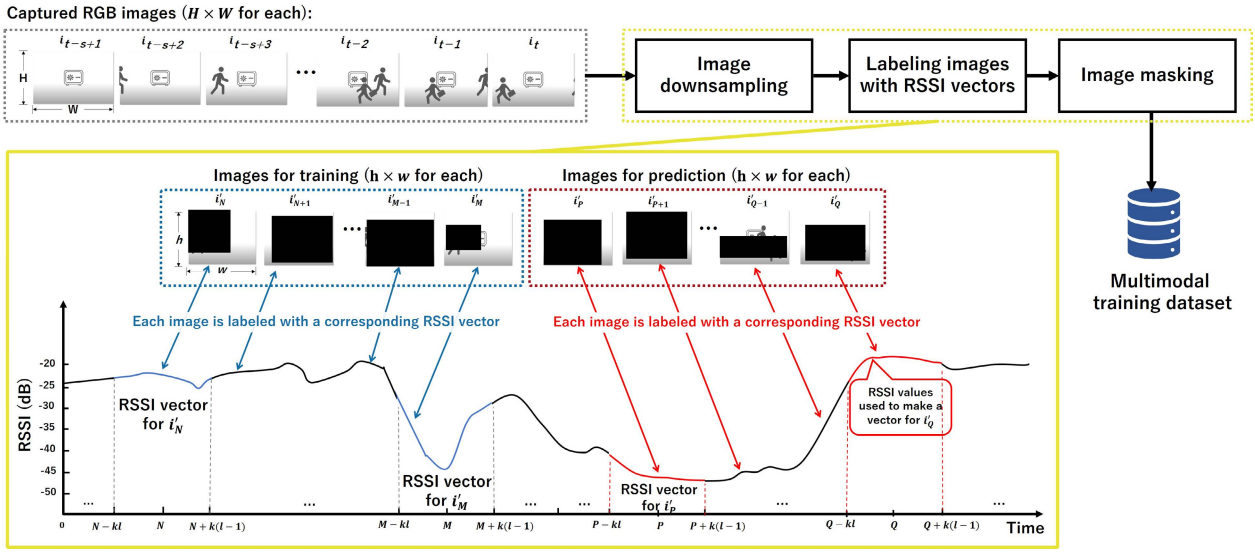

FIGURE 4. The workflow for generating multimodal training dataset.

D. MODEL ARCHITECTURE

The MMI model is a modified U-net architecture consisting of two encoders (an image encoder and an RSSI encoder) and one decoder. To highlight the contribution of RSSI to the inpainting performance, two MMI models containing one and four RSSI data input channels are constructed, respectively. The structure of the MMI model with four RSSI-input is depicted (see Fig. 5). The structure of the other model is identical to this one except the number of RSSI inputs.

The image encoder accepts a 64×64-pixel defective RGB image as input. Subsequently, four downsampling modules are used. Each downsampling module sequentially contains a 2D convolution layer, a batch normalization layer (only for the last three modules), and a leaky ReLU layer. The data is transmitted through all four downsampling modules, yielding a $4 \times 4 \times 256$ output tensor.

An RSSI encoder is designed to extract spatial features from an RSSI vector. RSSI data obtained from each of the four APs are transmitted to its four input layers, and subsequently fused in a concatenate layer to obtain a $1 \times 4L$ tensor. The tensor first enters a batch normalization layer, then passes through a structure comprising a dense layer, a batch normalization layer, and a ReLU layer thrice successively. Finally, the tensor shape is transformed into $4 \times 4 \times 8$ using a reshaping layer. At the bottleneck of the MMI model, a concatenate layer is employed to fuse the $4 \times 4 \times 256$ tensor obtained from the image encoder with the $4 \times 4 \times 8$ tensor obtained from the RSSI encoder, which is then fed into the decoder.

The decoder contains four upsampling modules. Each of the first three modules consists of a transposed 2D convolution layer, a batch normalization layer, and a ReLU layer. After passing through these three modules, the tensor is transmitted to a transposed 2D convolution layer with an activation function of tanh and a filter number of three. Skip connections are used to link the image encoder and decoder at different levels of spatial feature abstraction. Finally, the decoder outputs a reconstructed image with the size of $64 \times 64 \times 3$. In the entire model, only convolution and transposed convolution layers with stride of two and filter size of three are used. Adam is used as the optimizer for the MMI model, the mean squared error (MSE) is used as the loss function.

IV. EVALUATION

A. EXPERIMENTAL CONFIGURATION

The feasibility of implementing multimodal image inpainting using RF-Inpainter is evaluated in a typical indoor environment. The experiment comprises two phases—data acquisition and model training. The facile application of RF-Inpainter to various daily scenarios is demonstrated by acquiring experimental data using off-the-shelf devices (e.g., WiFi routers and laptops). In particular, four WZR-HP-AG300H APs manufactured by Buffalo are used. These APs exhibit a maximum data transmission rate of 54 Mbps and transmit beacon frames in the 5 GHz band at 100 ms intervals. The beacon frames are captured using the laptop to measure the RSSI values obtained from different APs. The APs are placed on four shelves located on a straight line along the wall—each shelf is approximately one meter high, and the distance between two adjacent shelves is also nearly one meter. The distance between the RF sensors and APs is approximately 3–4 m. Both RGB images and corresponding RSSI are collected simultaneously on the laptop at a frame rate of 10 fps for approximately 10 min. Fig. 6 depicts the experimental configuration, and Fig. 7 and 8 illustrate the indoor experimental environment from the perspectives of Camera 1 and Camera 2, respectively.

The simplest possible indoor application scenario is utilized, comprising a single moving pedestrian in each camera's view in addition to the background. The trajectory of this pedestrian is taken to lie between the APs and the RF sensor (see Fig. 6). The trajectory is 8 m long, with vertical distances of approximately 1.5 m from the APs and approximately

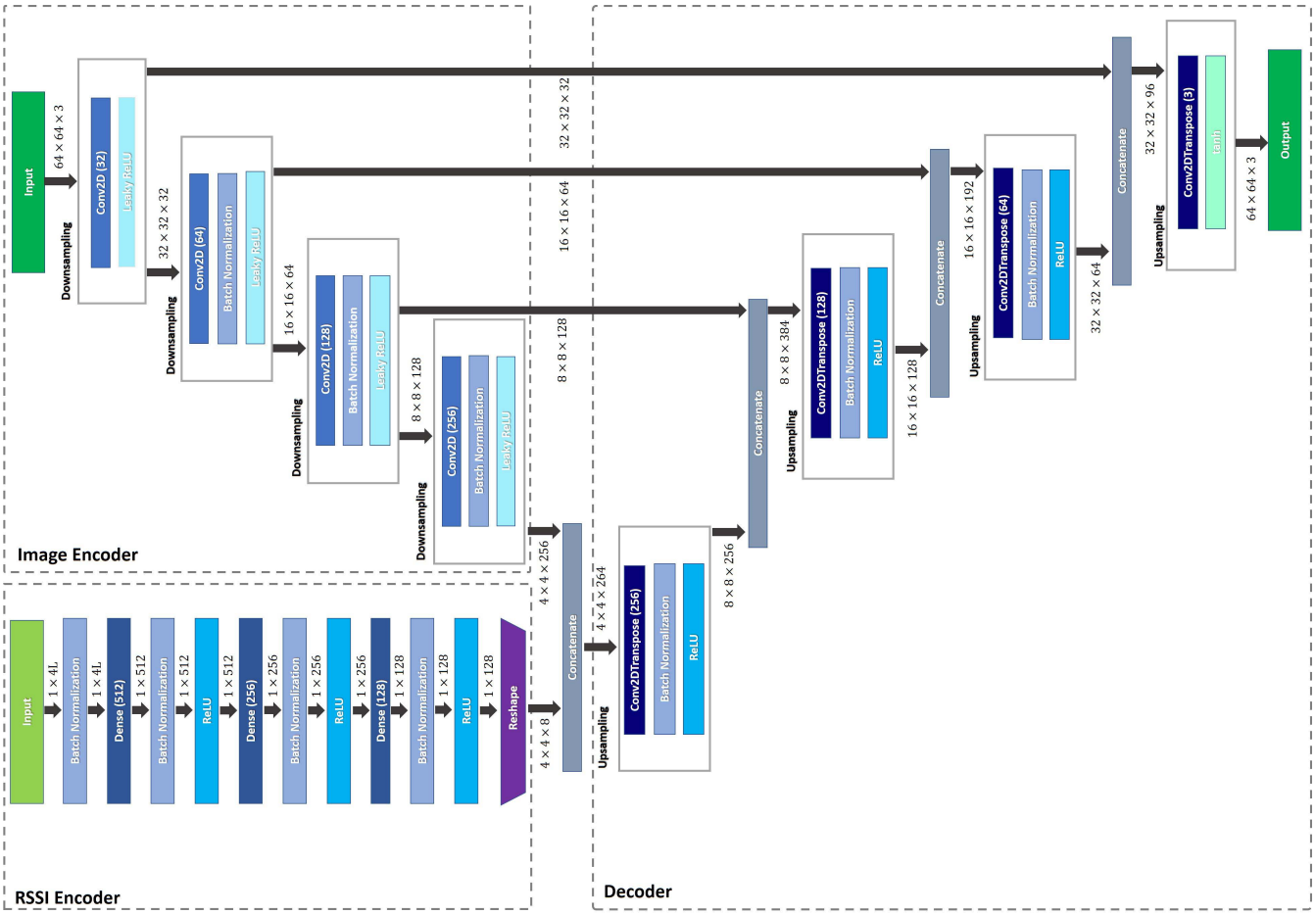

FIGURE 5. The architecture of the MMI model with four RSSI-input.

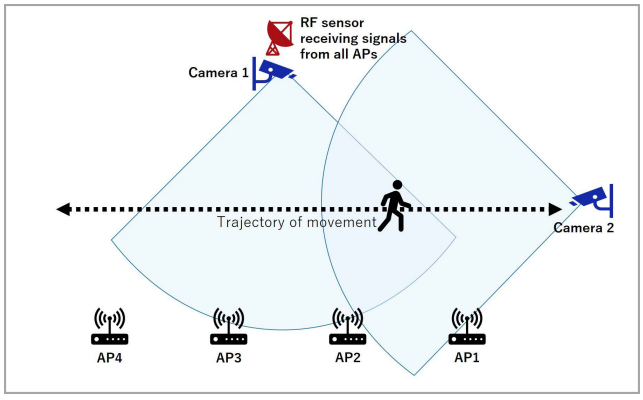

FIGURE 6. The experimental configuration.

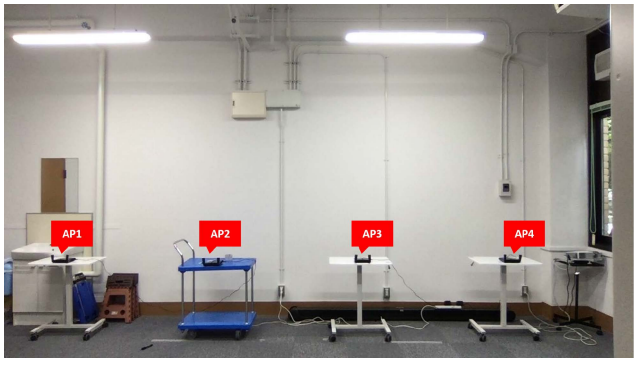

FIGURE 7. A snapshot of the experimental environment from the perspective of Camera 1.

2.5 m from the RF sensor. The entire walk is observed to last approximately 20 s on average. To imitate the passage of pedestrians through the cameras' views from different directions, the pedestrian moves back and forth along the trajectory. During this process, Camera 1 captures the sides of the pedestrian, while Camera 2 captures images depicting the back and front of the pedestrian.

The movement of the pedestrian is observed to affect the RSSI values measured by the RF sensor by varying degrees. When the pedestrian walks close to an AP, most WiFi signals emitted by the AP are reflected by the pedestrian—thus, the RSSI obtained from that particular AP is reduced. As a result, the RSSI values fluctuate, and the degree of this fluctuation affects inpainting results significantly—the more extensive the fluctuation, the richer the environmental information contained in the RSSI, and consequently, the more beneficial it is to the production of a clear image. To emphasize RSSI fluctuation, the pedestrian is asked to hold a white board

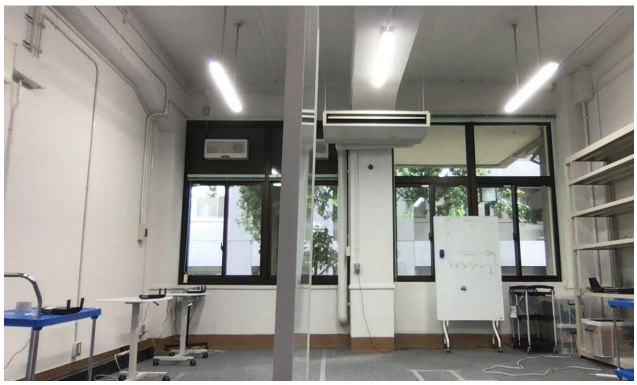

FIGURE 8. A snapshot of the experimental environment from the perspective of Camera 2.

while walking, which increases the proportion of reflected WiFi signals.

In the aforementioned experimental configuration, 5745 and 5740 clear images with a resolution of 1280 × 720 pixels are captured using Camera 1 and 2, respectively. The image size is reduced to 64 × 64 pixels after downsampling. 6029, 6036, 6033, and 6032 RSSI values are received by the RF sensors from AP1, AP2, AP3, and AP4, respectively. Subsequently, these collected data are used to produce a multimodal dataset following the method described in Section III-C, assuming the length of the RSSI vector to be 200.

Next, the multimodal dataset, D, is divided into training and validation datasets. To avoid data leakage, which produces models with high training performance but poor prediction performance, time-based data splitting with a splitting ratio of 80% is utilized.

Finally, the models with split datasets are trained for 30 epochs and validated. Training and validation are conducted using Google Colab on a Tesla T4 GPU.

B. METRICS FOR THE EVALUATION OF INPAINTED IMAGES

The two most commonly used objective image quality assessment (IQA) metrics for complete reference images are sued to measure the quality of the reconstructed images quantitatively.

- Peak Signal-to-Noise Ratio (PSNR)
  PSNR is a well-known error-based objective IQA metric. It represents the average of the squares of the "errors" between the original image and the degraded image. PSNR is positively correlated to the similarity between the reconstructed image and the original image, i.e., the quality of inpainting performance.
- Structural Similarity Index (SSIM)
  PSNR is not highly indicative of perceived similarity during image comparison, and SSIM aims to address this shortcoming. SSIM is a structural similarity-based metric used to measure the similarity between a reference image and a degraded image by taking texture into account. SSIM takes a value between 0 and 1, which is positively correlated with the quality of the reconstructed image, i.e., the accuracy of the image inpainting method.

Additionally, the efficiency of RF-Inpainter is quantitatively evaluated—to this end, the mean inference time of inpainting using the MMI model is calculated. The inference time presents the duration required for a forward propagation process that, given an input, obtains the output. In this experiment, and the mean inference time is the average time required by a trained model to restore a test image. The mean inference time is inversely correlated with the efficiency of the inpainting method.

C. BASELINE METHODS

Single-modal image inpainting methods, i.e., image-only inpainting and RSSI-only inpainting methods, are considered as baselines. Specifically, single-modal inpainting models are constructed using images or RSSI as input.

The structure of the image-only model is identical to the U-Net structure in the MMI model, with the exception that the RSSI encoder and concatenate layer used to fuse image and RSSI information are removed.

The RSSI-only inpainting model is based on a modified auto-encoder architecture. To highlight the fact that increasing RSSI information to a certain extent improves the quality of images, as in the MMI models, RSSI-only inpainting models with only one RSSI channel and four RSSI channels are constructed. These models are denoted by RSSI-only inpainting (w/ single AP) and RSSI-only inpainting (w/ 4 APs), respectively. The encoder of the RSSI-only inpainting models follows the RSSI encoder in the MMI model, and only the output size of its reshaping layer is changed to $8 \times 8 \times 2$. The output tensor of the RSSI encoder passes through three successive upsampling structures in the decoder, each containing one 2D convolutional layer and one 2D upsampling layer. All 2D convolutional layers contain a convolutional kernel with a size of three, stride of one, and ReLU as the activation function. The 2D upsampling layers contain upsampling factors of 2×2 for rows and columns. Finally, the tensor is transmitted through two 2D convolutional layers and a sigmoid activation layer, resulting in an output of size $64 \times 64 \times 3$.

The inpainting performances of the aforementioned baseline methods are compared with those of RF-Inpainter when four APs are available and when only a single AP is available, that i.e., with those of RF-Inpainter (w/ 4 APs) and RF-Inpainter (w/ single AP).

D. IMAGE OCCLUSION SCENARIOS

Three scenarios are considered based on the following occlusion patterns—horizontal, vertical, and random image occlusions.

In the horizontal occlusion scenario, nearly 70% of the lower area of all images is occluded, ensuring that the portrait is completely obscured in the images captured by Camera 1. The masking method is also used for images captured

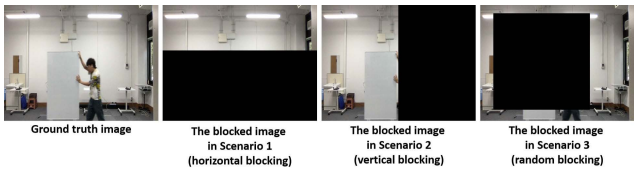

FIGURE 9. An example of occlusion of Camera 1's images in each scenario.

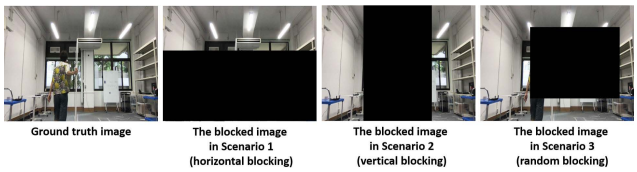

FIGURE 10. An example of occlusion of Camera 2's images in each scenario.

by Camera 2. From Camera 1's perspective, the pedestrian always moves back and forth horizontally along the same line—so information about the pedestrian exists on a fixed horizontal band. In the images captured by Camera 2, the pedestrian also walks linearly, but his distance from the camera keeps varying. Overall, the portrait changes dynamically in size and is always confined to a fixed vertical rectangular area. Thus, adding a horizontal block can remove all information about the pedestrian from Camera 1, but not from Camera 2 images, as parts of his body, such as the head and hands, are sometimes visible at the top of some images.

In the vertical occlusion scenario, 50% of the middle region of each Camera 2 image and 50% of the right region of all Camera 1 images are occluded. This method allows the complete occlusion of portraits in Camera 2 images, and pedestrians in all Camera 1 images are left unmasked with approximately half probability.

In the random occlusion scenario, the minimum size of the random blocking object on each image is taken to be 40×40 to ensure that most of the pedestrians in the images are occluded, regardless of whether the image comes from Camera 1 or 2.

E. EVALUATION OF RESULTS

Fig. 11 depicts a sample of the experimental results. The RF-Inpainter (w/ 4 APs) is observed to recover clear and complete portraits in all scenarios using multimodal information, irrespective of the occlusion of the portrait information in the input image. Conversely, RF-Inpainter (w/ single AP) and Image-only inpainting only recover portraits when residual portraits are present in the input images (i.e., in the random occlusion case), and these portraits are not as clear and complete as those reconstructed by RF-Inpainter (w/ 4 APs). In addition, RSSI-only inpainting (w/ 4 APs) recovers portraits that closely resemble the original ones. In contrast, RSSI-only inpainting (w/ single AP) does not reconstruct any portrait. Visually, the output of the four-input RSSI-only inpainting model is comparable to that of the four-input MMI model in Scenario 1 and 2. These observations indicate that using both image and RSSI information significantly enhances the robustness and accuracy of image inpainting. Furthermore, RSSI data obtained from multiple APs are required to improve inpainting performance.

Table 1 lists the objective assessment metrics (i.e., mean PSNR and mean SSIM) for all inpainting methods. In image occlusion scenarios, especially in the first two cases where portrait information is scarce, the accuracy of RF-Inpainter (w/ 4 APs) is the highest, followed by that of RF-Inpainter (w/ single AP). Image-only inpainting exhibits the worst accuracy. The maximum difference between its mean PSNR and that of RF-Inpainter (w/ single AP) is approximately 8 dB, and the maximum discrepancy in SSIM is approximately 12%. These observations reinforce the conclusion that the simultaneous use of both types of information improves image inpainting performance considerably.

Moreover, in the random blocking scenario, the accuracy of all three models is observed to be significantly improved compared to the first two scenarios, and all values are approximately comparable. Randomly generated occlusion often does not mask all portraits in an image, leading to leakage of portrait information, which aids inpainting. The main structure of MMI models and the Image-only inpainting model is U-Net, which is a robust neural network capable of reconstructing images to match the original images closely, even based on a small amount of information. As a result, the contribution of image information to image inpainting is significantly higher than that of RSSI—the image information almost completely determines the accuracy of the reconstructed images in this case. Therefore, RF-Inpainter and Image-only inpainting methods always yield similar results. Moreover, even when the percentage of the occluded area in the image is increased (e.g., the minimum size of the occluded area is 60×60), identical results are acquired, owing to the inclusion of cases where the randomly generated occluded area fails to block the pedestrian completely.

The results obtained in the RSSI-only scenarios demonstrate that images of satisfactory quality can be obtained using only RSSI, although the results are not as good as those obtained using RF-Inpainter. The results also suggest that appropriately increasing the quantity of RSSI data improves imaging performance.

Additionally, the accuracy of the reconstructed Camera 1 images is observed to be significantly higher than that of the inpainted Camera 2 images, which may be attributed to the positional relationship of APs, cameras, RF anchors, and the movement trajectory of the pedestrian.

From the perspective of efficiency, the mean inference time of each model is similar in all cases and is less than 1 ms. Further, the maximum inference time is only 4.097 ms, which enables the real-time restoration of images and facilitates the practical application of RF-Inpainter.

F. THE EFFECT OF RSSI VECTOR LENGTH

The RSSI vector length, L, affects the quality of the reconstructed images. To ensure that a vector contains sufficient RSSI values to capture the spatial environment completely, L

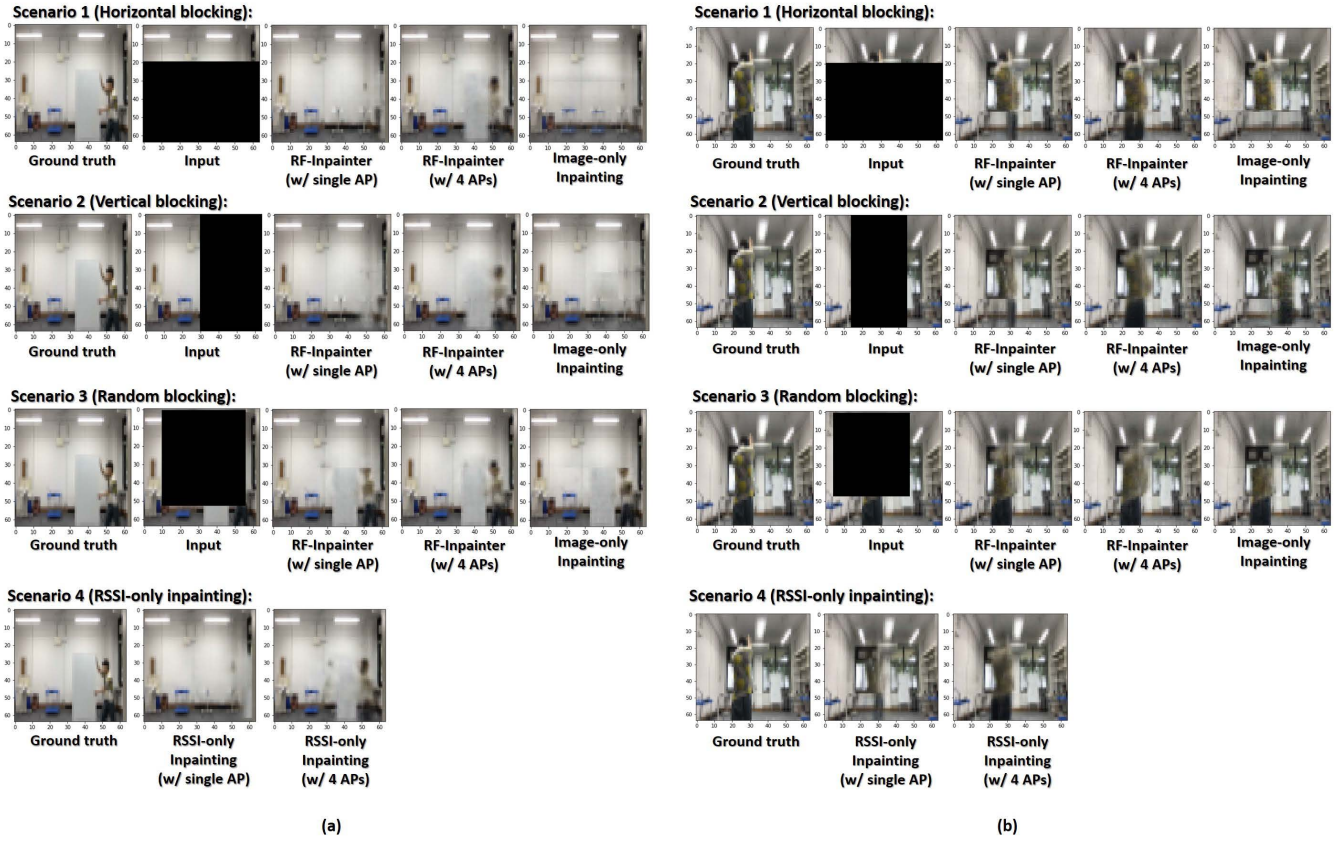

FIGURE 11. Sample inpainting results in each scenario: (a). Results for Camera 1 images; (b). Results for Camera 2 images.

TABLE 1. Objective assessment metrics for each inpainting method corresponding to Camera 1 images.

Inpainting method	In Camera 1 Dataset		In Camera 2 Dataset	
	Mean PSNR in Scenarios 1–3 or Scenario 4 (dB)	Mean SSIM in Scenarios 1–3 or Scenario 4	Mean PSNR in Scenarios 1–3 or Scenario 4 (dB)	Mean SSIM in Scenarios 1–3 or Scenario 4
RF-Inpainter (w/ single AP)	27.73 / 30.77 / 33.31	0.936 / 0.962 / 0.974	24.54 / 23.65 / **27.82**	0.918 / 0.914 / **0.935**
RF-Inpainter (w/ 4 APs)	**29.92 / 32.92 / 35.02**	**0.955 / 0.976 / 0.979**	**25.08 / 25.50** / 27.57	**0.922 / 0.923** / 0.932
Image-only Inpainting	21.64 / 26.90 / 33.79	0.833 / 0.943 / 0.974	22.68 / 18.50 / 26.92	0.899 / 0.824 / 0.933
RSSI-only Inpainting (w/ single AP)	26.897	0.927	22.831	0.859
RSSI-only Inpainting (w/ 4 APs)	28.935	0.949	23.264	0.873

must not be too small. However, if L is too large, interference information is increased, which impairs the inpainting performance. Therefore, optimizing the value of L is essential.

Fig. 12 depicts the images reconstructed by RF-Inpainter (w/ 4 APs) with RSSI vector lengths of 150, 250, and 550. When the number of RSSI values in the vector is insufficient ($L = 150$), the reconstructed human image is blurred, as depicted in Fig. 12. a. In contrast, when too many RSSI values are included in the vector ($L = 550$), the result depicted in Fig. 12. c is obtained—the portrait in the reconstructed image does not match the original image. However, when the number of RSSI values included in the vector is moderate ($L = 250$), an accurate and precise result is obtained, as depicted in Fig. 12. b. Thus, it is reasonable to believe that there might be an optimum value of L (i.e., m) between 150 and 550 that maximizes inpainting accuracy.

Similar conclusions can be drawn from Fig. 13 and 14 that capture the variations of the mean PSNR and mean SSIM of the reconstructed images as a functions of L, separately. A peak is observed in the middle of each curve in each

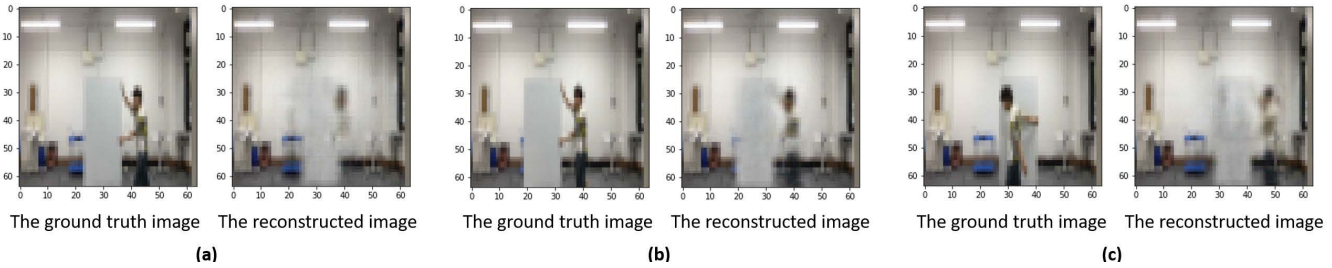

FIGURE 12. Image inpainting results corresponding to different RSSI vector lengths. (a) $L = 150$. (b) $L = 250$. (c) $L = 550$.

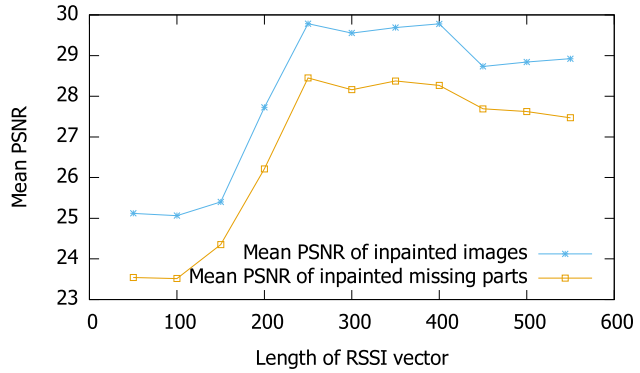

FIGURE 13. Dependence of mean PSNR on the length of RSSI vector.

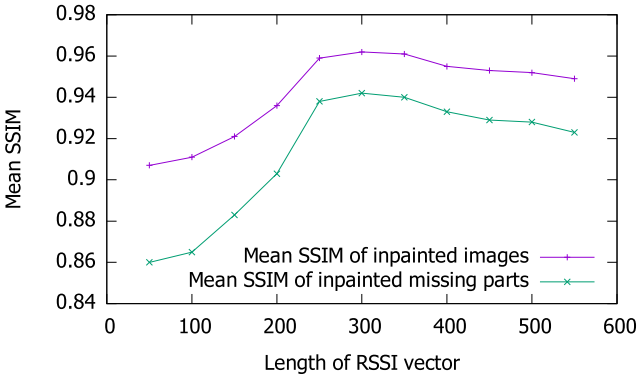

FIGURE 14. Dependence of the mean SSIM on the length of RSSI vector.

graph, indicating the existence of maximum values for both PSNR and SSIM. For mean PSNR and mean SSIM, the peaks correspond to the range of L between 200 and 300, and between 250 and 350, respectively.

However, the accurate determination of m in a particular environment is a daunting task owing to a great number of factors affecting m, such as the positional relationship between APs and the RF sensor, the density of human traffic in the space, and the size of the indoor space. In addition, most of these factors change dynamically over time, causing m to change constantly. Thus, we reserve the precise determination of m in real time as a topic for future research.

V. CONCLUSION

This article propose RF-Inpainter—a novel multimodal image inpainting method that integrates visual and wireless information. The underlying architecture of RF-Inpainter is a deep neural network called MMI model, which restores complete and clear images by fusing temporally corresponding defective RGB images and RSSI vectors. The feasibility and advantages of inpainting using RF-Inpainter are illustrated by evaluating the inpainting performances of two MMI models and three baseline models in a typical indoor environment using experimentally obtained datasets. The results reveal that the fusion of RF information improves image quality significantly in most scenarios, with maximum improvements in mean PSNR and mean SSIM of 36.4% and 14.6%, respectively. Moreover, the mean inference time of the MMI model is lower than 1 ms, which indicates that RF-Inpainter enables real-time restoration of defective images.

We expect several directions of future research to emerge regarding this work as a baseline. Firstly, the evaluation of inpainting performance of RF-Inpainter by applying it to computer vision applications, such as object recognition and moving path prediction, rather than solely based on metrics such as mean PSNR and mean SSIM, may be desirable. Secondly, we intend to explore optimizing the performance of both wireless communication and image inpainting in future work. In other words, we expect to maximize the coverage and throughput of a wireless communication network by determining the optimal location of access points, RF sensors and cameras, while ensuring good sensing capabilities of that network.

ACKNOWLEDGMENT

The authors would like to thank Dr. Ryo Yonetani for the insightful comments and helpful discussions.

REFERENCES

[1] A. R. Widya, Y. Monno, M. Okutomi, S. Suzuki, T. Gotoda, and K. Miki, "Whole stomach 3D reconstruction and frame localization from monocular endoscope video," *IEEE J. Transl. Eng. Health Med.*, vol. 7, pp. 1–10, 2019.

[2] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii, "InLoc: Indoor visual localization with dense matching and view synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7199–7209.

[3] T. Sattler, A. Torii, J. Sivic, M. Pollefeys, H. Taira, M. Okutomi, and T. Pajdla, "Are large-scale 3D models really necessary for accurate visual localization?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1637–1646.

[4] J. Jam, C. Kendrick, K. Walker, V. Drouard, J. G.-S. Hsu, and M. H. Yap, "A comprehensive review of past and present image inpainting methods," *Comput. Vis. Image Understand.*, vol. 203, Feb. 2021, Art. no. 103147.

[5] X. Zhuge, "Short-range ultra-wideband imaging with multiple-input multiple-output arrays," Ph.D. thesis, Dept. Telecommun., Delft Univ. Technol., Delft, The, Netherlands, 2010. [Online]. Available: http://resolver.tudelft.nl/uuid:5a7ce119-6ed2-420b-9a5a-200896fb3445
[6] W. Zhong, K. He, and L. Li, "Through-the-wall imaging using WiFi signals," *J. Eng.*, vol. 2019, no. 20, pp. 6940–6942, 2019.
[7] S. Vakalis, L. Gong, and J. A. Nanzer, "Imaging with WiFi," *IEEE Access*, vol. 7, pp. 28616–28624, 2019.
[8] S. Fowler, G. G. Baravdish, and G. Baravdish, "3D imaging of sparse wireless signal reconstructions via machine learning," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2020, pp. 1–6.
[9] S. Zhou, L. Guo, Z. Lu, X. Wen, W. Zheng, and Y. Wang, "Subject-independent human pose image construction with commodity Wi-Fi," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2021, pp. 1–6.
[10] C. Oliver, "Synthetic-aperture radar imaging," *J. Phys. D, Appl. Phys.*, vol. 22, no. 7, p. 871, 1989.
[11] L. Ferro-Famil and E. Pottier, "SAR imaging using coherent modes of diversity: SAR polarimetry, interferometry and tomography," in *Microwave Remote Sensing of Land Surface*. U.K.: Elsevier, 2016, pp. 67–147. [Online]. Available: https://www.sciencedirect.com/book/9781785481598/microwave-remote-sensing-of-land-surfaces
[12] S. Kato, T. Fukushima, T. Murakami, H. Abeysekera, Y. Iwasaki, T. Fujihashi, T. Watanabe, and S. Saruwatari, "CSI2Image: Image reconstruction from channel state information using generative adversarial networks," *IEEE Access*, vol. 9, pp. 47154–47168, 2021.
[13] A. Dubey, P. Sood, J. Santos, D. Ma, C.-Y. Chiu, and R. Murch, "An enhanced approach to imaging the indoor environment using WiFi RSSI measurements," *IEEE Trans. Veh. Technol.*, vol. 70, no. 9, pp. 8415–8430, Sep. 2021.
[14] M. H. Kefayati, V. Pourahmadi, and H. Aghaeinia, "Wi2 Vi: Generating video frames from WiFi CSI samples," *IEEE Sensors J.*, vol. 20, no. 19, pp. 11463–11473, Oct. 2020.
[15] L. Guo, Z. Lu, X. Wen, S. Zhou, and Z. Han, "From signal to image: Capturing fine-grained human poses with commodity Wi-Fi," *IEEE Commun. Lett.*, vol. 24, no. 4, pp. 802–806, Apr. 2020.
[16] N. M. F. Salem, "A survey on various image inpainting techniques," *Future Eng. J.*, vol. 2, no. 2, p. 1, 2021.
[17] Y. Ma, X. Liu, S. Bai, L. Wang, A. Liu, D. Tao, and E. R. Hancock, "Regionwise generative adversarial image inpainting for large missing areas," *IEEE Trans. Cybern.*, early access, Aug. 17, 2022, doi: 10.1109/TCYB.2022.3194149.
[18] J. Qin, H. Bai, and Y. Zhao, "Face inpainting network for large missing regions based on weighted facial similarity," *Neurocomputing*, vol. 386, pp. 54–62, Apr. 2020.
[19] J. Li, N. Wang, L. Zhang, B. Du, and D. Tao, "Recurrent feature reasoning for image inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7760–7768.
[20] S. S. Ahmed, A. Schiessl, and L.-P. Schmidt, "A novel fully electronic active real-time imager based on a planar multistatic sparse array," *IEEE Trans. Microw. Theory Techn.*, vol. 59, no. 12, pp. 3567–3576, Dec. 2011.
[21] P. M. Holl and F. Reinhard, "Holography of Wi-Fi radiation," *Phys. Rev. Lett.*, vol. 118, no. 18, May 2017, Art. no. 183901.
[22] J. Hunt, T. Driscoll, A. Mrozack, G. Lipworth, M. Reynolds, D. Brady, and D. R. Smith, "Metamaterial apertures for computational imaging," *Science*, vol. 339, no. 6117, pp. 310–313, Jan. 2013.
[23] L. Yujiri, M. Shoucri, and P. Moffa, "Passive millimeter wave imaging," *IEEE Microw. Mag.*, vol. 4, no. 3, pp. 39–50, Sep. 2003.
[24] T. Nishio, Y. Koda, J. Park, M. Bennis, and K. Doppler, "When wireless communications meet computer vision in beyond 5G," *IEEE Commun. Standards Mag.*, vol. 5, no. 2, pp. 76–83, Jun. 2021.
[25] T. Nishio, H. Okamoto, K. Nakashima, Y. Koda, K. Yamamoto, M. Morikura, Y. Asai, and R. Miyatake, "Proactive received power prediction using machine learning and depth images for mmWave networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 11, pp. 2413–2427, Nov. 2019.
[26] Y. Koda, K. Nakashima, K. Yamamoto, T. Nishio, and M. Morikura, "Handover management for mmWave networks with proactive performance prediction using camera images and deep reinforcement learning," *IEEE Trans. Cognit. Commun. Netw.*, vol. 6, no. 2, pp. 802–816, Jun. 2020.
[27] J. Jiao, F. Li, W. Tang, Z. Deng, and J. Cao, "A hybrid fusion of wireless signals and RGB image for indoor positioning," *Int. J. Distrib. Sens. Netw.*, vol. 14, no. 2, pp. 1–11, 2018.
[28] D. Zhu, H. Sun, and D. Wu, "Fusion of wireless signal and computer vision for identification and tracking," in *Proc. 28th Int. Conf. Telecommun. (ICT)*, Jun. 2021, pp. 1–7.
[29] Z. Yang, Z. Zhou, and Y. Liu, "From RSSI to CSI: Indoor localization via channel response," *ACM Comput. Surv.*, vol. 46, no. 2, pp. 1–32, 2013.

CHENG CHEN received the B.E. degree in information and communications engineering from the National University of Defence Technology, in 2019. He is currently pursuing the M.I. degree with the School of Engineering, Tokyo Institute of Technology.

TAKAYUKI NISHIO (Senior Member, IEEE) received the B.E. degree in electrical and electronic engineering and the master's and Ph.D. degrees in informatics from Kyoto University, in 2010, 2012, and 2013, respectively. He was an Assistant Professor in communications and computer engineering at the Graduate School of Informatics, Kyoto University, from 2013 to 2020. From 2016 to 2017, he was a Visiting Researcher at the Wireless Information Network Laboratory (WINLAB), Rutgers University, USA. Since 2020, he has been an Associate Professor at the School of Engineering, Tokyo Institute of Technology, Japan, and the Wireless Information Network Laboratory (WINLAB), Rutgers University. His current research interests include machine learning-based network control, machine learning in wireless networks, vision-aided wireless communications, and heterogeneous resource management.

MEHDI BENNIS (Fellow, IEEE) is currently a Professor with the Centre for Wireless Communications, University of Oulu, Finland, an Academy of Finland Research Fellow, and the Head of the Intelligent Connectivity and Networks/Systems Group (ICON). He has published over 200 research papers in international conferences, journals, and book chapters. His research interests include radio-resource management, heterogeneous networks, game theory, and distributed machine learning in 5G networks and beyond. He has been a recipient of several prestigious awards, including the 2015 Fred W. Ellersick Prize from the IEEE Communications Society, the 2016 Best Tutorial Prize from the IEEE Communications Society, the 2017 EURASIP Best Paper Award for the *Journal of Wireless Communications and Networking*, the University of Oulu Award for Research, the 2019 IEEE ComSoc Radio Communications Committee Early Achievement Award, and the 2020 Clarviate Highly Cited Researcher from the Web of Science. He is also an Editor of IEEE TRANSACTIONS ON COMMUNICATIONS and the Specialty Chief Editor for *Data Science for Communications* and the *Frontiers in Communications and Networks* journal.

JIHONG PARK (Senior Member, IEEE) received the B.S. and Ph.D. degrees from Yonsei University, South Korea. He is currently a Lecturer at the School of IT, Deakin University, Australia. His research interests include ultra-dense ultra-reliable mmWave system design, distributed learning control ledger technologies, and their applications in beyond-5G/6G communication systems. He served as a Conference Workshop Program Committee Member for IEEE GLOBECOM, ICC, and WCNC, and for NeurIPS, ICML, and IJCAI. He is also an Associate Editor of Frontiers in *Data Science for Communications* and a Review Editor of frontiers in *Aerial and Space Networks*.