Locating Multiple RFID Tags with Swin Transformer-based RF Hologram Tensor Filtering

[†]Xiangyu Wang, [‡]Jian Zhang, [†]Shiwen Mao, [§]Senthilkumar CG Periaswamy, and [§]Justin Patton
[†]Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849-5201
[‡]Department of Electrical and Computer Engineering, Kennesaw State University, Kennesaw, GA 30144
[§]RFID Lab, Auburn University, Auburn, AL 36849
Email: xzw0042@auburn.edu, {jianzhang, smao}@ieee.org, {szc0089, jbp0033}@auburn.edu

2022 IEEE 96th Vehicular Technology Conference (VTC2022-Fall) | 978-1-6654-5468-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/VTC2022-Fall57202.2022.10013016

***Abstract*—In this paper, we present a Swin Transformer based indoor localization framework that employs RF hologram tensors to locate multiple ultra-high frequency (UHF) passive Radio-frequency identification (RFID) tags. The RF hologram tensor captures the strong relationship between RFID measurements and spatial location, and helps to improve the robustness of the system in dynamic environments. We develop a Swin Transformer-based hologram filter network to clean the fake peaks in hologram tensors caused by multipath propagation and phase wrapping, exploring the spatial relationship between tags. In contrast to fingerprinting-based localization systems that use deep networks as classifier, the proposed network treats localization as a regression problem. An intuitive peak finding algorithm is introduced for location estimation using the sanitized hologram tensors. We prototype the proposed system using commodity RFID devices and conduct extensive experiments to evaluate its performance.**

***Index Terms*—Radio-frequency identification (RFID), Indoor localization, Swin Transformer, self-supervised learning.**

I. Introduction

Indoor localization has remained a hot research topic over the years, as it plays a critical role in solving position-related problems such as gesture recognition and human pose estimation [1]. Recently, researchers bring deep networks into indoor localization systems cooperating with the fingerprinting method. However, several inherent problems of fingerprinting based localization systems are still open. Collecting fingerprints in a large area would be laborious and time-consuming. The minimum error of the fingerprinting-based localization relies on the granularity of the stored fingerprints.

In this paper, Radio frequency (RF) hologram tensors are created using phase readings from received RFID response signals as input to a deep learning model for locating multiple tags, as in our previous work MulTLoc [2]. We implement a Swin Transformer-based hologram filter network for locating multiple RFID tags with the Hologram tensors. Self-supervised pre-training is leveraged to extract the general features from the hologram tensors for improved localization performance. Location estimation could be accomplished with the sanitized hologram tensors directly with an intuitive peak detection algorithm. The experimental results demonstrate the superior performance of the proposed system.

II. System Design

A. Architecture Overview

Fig. 1 depicts the architecture of the proposed system. As in our previous work MulTLoc [2], an RFID system collaborates with a vision-based sensor to generate the hologram tensors and the accompanying ground truth tensors in order to train the networks for location estimation. The Robot Operating System (ROS) is utilized to synchronize and unify the data acquired from diverse hardware. The noisy hologram tensors are sanitized with a Swin Transformer based network. Tag locations are estimated by a peak detection algorithm.

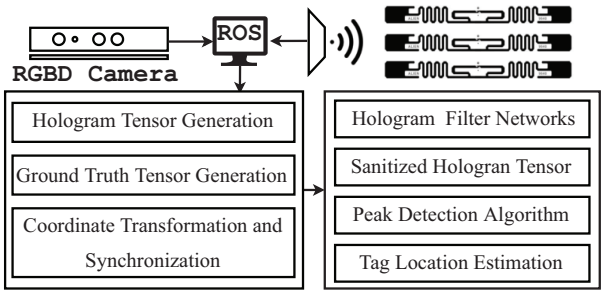

Fig. 1. The MulTLoc system architecture.

B. Network Architecture of the Hologram Filter Network

A Swin Transformer-based network is leveraged to sanitize the noisy hologram tensors and ensure that the output tensor has the same size as the input noisy hologram tensor. The architecture of the Swin Transformer based hologram filter network is depicted in Fig. 2(a). The network is a 3D variation of the U-Net [3] with a Swin Transformer backbone. To be fed into the Swin Transformer blocks, the input tensor is first separated into non-overlapping 3D tokens. The patch merging layer reduces the feature size by a factor of two in each stage of the Swin Transformer backbone. The output of each stage will not only be sent on to the next stage, but also be fed into the DCNN based decoders to recreate the filtered hologram tensor. The filtered hologram tensor is obtained directly from the convolutional decoder of the top layer. A function consisting of L_1 loss and MS-SSIM loss is used as the loss function to supervise the training [4]. A hyper-parameter α,which is set as 0.6, is utilized for balancing two parts in the loss function.

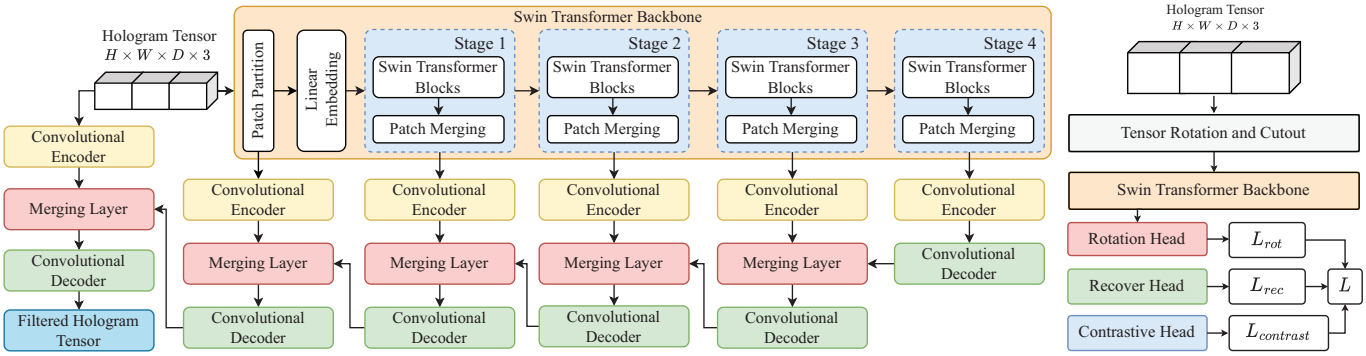

Fig. 2. (a) Architecture of the Swin Transformer based hologram filter network. (b) Architecture of the self-supervised training.

C. Self-supervised Pre-training of Swin Transformers

Self-supervised pre-training, as shown in Fig. 2(b), is leveraged to extract general features in the hologram tensors. We adopted three pretext loss for learning a good data representation, which are inspired by the prior work for medical image analysis [5]. The hologram tensor is rotated in six angle classes, including $90°, -90°$ along with X-, Y-, and Z-axis, to generate sub-volumes randomly. A cross-entropy loss, L_{rot}, is leveraged to predict rotation angles. A tensor recovery task is the second part of the self-supervised training. To recover the mask-out pixels in the sub-volumes, the MSE loss, L_{rec} is leveraged to measure the difference between the original sub-volume and the recovered sub-volume. A simple instance discrimination task is utilized as the third pretext task. Two correlated sub-volumes of an input hologram tensor are generated with rotation and cutout. For a minibatch, only the feature representation from the same input tensor is treated as positive pair, while the representation from the rest tensors is the negative example. InfoNCE [6], $L_{contrast}$, is utilized as the contrastive loss function. Thus, the self-supervised training is conducted with the loss function, $L = L_{rot} + L_{rec} + L_{contrast}$. The hologram filter network is fine-tuned based on the pre-trained Swin Transformer backbone.

III. Experimental Study

To evaluate the performance of the proposed system, we built a prototype using a Zebra FX9600 reader and eight Zebra AN720 antennas. Three UPM Raflatac Frog 3D tags are utilized as localization targets. In the experiment, we assess the performance of the proposed framework by concurrently localizing the tags attached to the shoulders and neck of a subject. A Kinect V2 is leveraged to produce ground truth coordinates for network training and testing. The serviced area in our prototype covers an area of dimension 1.5m $\times$ 1.5m $\times$ 1.5m at a height of 0.5m above the ground. The grid size is set to 1cm. The estimation location is given by $\widehat{G} = \{G | f(\mathbf{S}_R, G) = \max(\mathbf{S}_R)\}$, where $f(\cdot)$ extracts the similarity value at the grid location G from the sanitized hologram tensor $\mathbf{S}_R$. Fig. 3 presents the cumulative distribution function (CDF) of localization errors, which exhibits the overall precision improvement brought by self-supervised learning. A mean error of $0.0961m$ is achieved with the self-supervised training, whereas the mean error is $0.1041m$ without self-supervised training. The improvement is 7.68%. The experiment demonstrates that self-supervised training contributes to the precision improvement in location estimation.

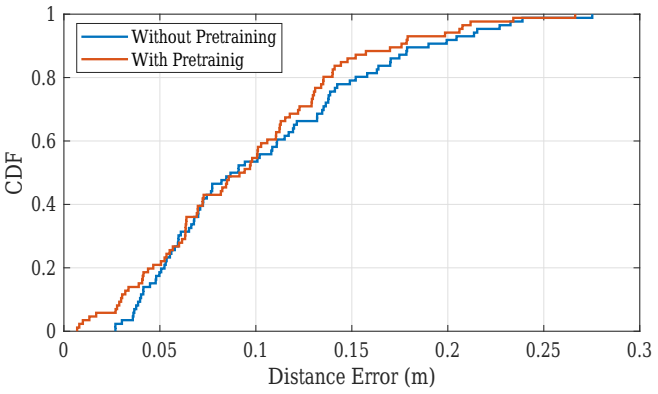

Fig. 3. CDFs of location estimation errors with and without the pretriaining of the Swin Transformer backone.

Acknowledgments

This work was supported in part by the NSF under Grants ECCS-1923163 and CNS-2107190.

References

[1] C. Yang, X. Wang, and S. Mao, "RFID-Pose: Vision-aided 3D human pose estimation with RFID," *IEEE Transactions on Reliability*, vol. 70, no. 3, pp. 1218–1231, Sept. 2021.

[2] X. Wang, J. Zhang, S. Mao, S. Periaswamy, and J. Patton, "MulTLoc: RF hologram tensor filteing and upscaling for indoor localization using multiple UHF passive RFID tags," in *Proc. ICCCN 2021*, Athens, Greece, July 2021.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI 2015*, Munich, Germany, Oct. 2015, pp. 234–241.

[4] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Transactions on Computational Imaging*, vol. 3, no. 1, pp. 47–57, Mar. 2017.

[5] Y. Tang, D. Yang, W. Li, H. R. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh, "Self-supervised pre-training of swin transformers for 3d medical image analysis," in *Proc. IEEE CVPR 2022*, New Orleans, LA, June 2022, pp. 20 730–20 740.

[6] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, July 2018. [Online]. Available: https://arxiv.org/abs/1807.03748