ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3315-6701-9/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICASSP55912.2026.11465014

BI-DIRECTIONAL ATTENTION FOR DUAL-BRANCH GENERATOR FOR CHANNEL EXTRAPOLATION AND HIGH-RESOLUTION SENSING

Jilong Du[1], *Qian Yang*[2], *Qianyi Huang*[1†], *Guochao Song*[3], *Xu Chen*[1], *Qian Zhang*[4]

[1] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2] Southern University of Science and Technology, Shenzhen, China
[3] China Academy of Information and Communications Technology, Beijing, China
[4] Hong Kong University of Science and Technology, Hong Kong, China

ABSTRACT

Integrated sensing and communication (ISAC) is a key technology in 5G-Advanced (5G-A) and 6G; however, current communication systems typically provide limited bandwidth ($<$100 MHz), which is unfavorable for sensing applications. To address this challenge, we propose a deep generative architecture for wideband channel state information (CSI) extrapolation: given narrowband CSI, it predicts the unknown wideband channel state. The key idea is a dual-branch generative architecture, a CNN branch and a transformer branch, where a two-stage, bi-directional cross-attention mechanism tightly couples the CNN's local features with the Transformer's global context modeling. Compared with the State-of-the-Art (SOTA) baseline, our model demonstrates significant performance gains across all key metrics, improving SNR by 4.5dB in the most challenging (dense-urban) scenarios. Crucially, when applied to a real-world Unmanned Aerial Vehicle (UAV) channel dataset, the extrapolated wideband CSI reduces the ranging error from 3.81m (20MHz observed CSI) to 0.56m (80MHz extrapolated CSI).

***Index Terms*—** ISAC, CSI extrapolation, dual-branch, bi-directional attention

1. INTRODUCTION

Integrated sensing and communication (ISAC) is a cornerstone technology for 5G-Advanced (5G-A) and 6G [1]. It is a well-known fact that sensing resolution scales inversely with signal bandwidth; yet cellular systems typically offer limited bandwidths (*e.g.*, 20 MHz in LTE and up to 100 MHz in 5G NR FR1). Under these constraints, the best achievable distance resolution is on the order of 3 m at 100 MHz (*i.e.*, speed of light/signal bandwidth), which is inadequate for certain applications, such as UAV tracking and formation [2].

This fundamental limitation motivates our central research question: can we extrapolate a wideband CSI from its readily available but narrowband counterpart, thereby synthesizing a large bandwidth required for high-resolution sensing? It falls within the typical channel prediction problem, with a representative case being the inference of downlink CSI from uplink CSI in FDD systems [3], where the uplink and donwlink bandwidth are typically symmetric. Existing approaches to these problems broadly fall into two categories. **Model-based methods** [4, 5] reconstruct the channel from a sparse set of estimated physical parameters (e.g., path delays and attenuations). The underlying propagation dynamics are far more complex over a wide bandwidth, causing the effective path parameters to vary across sub-bands and thus invalidating a single, fixed model [6]. **Data-driven methods** [3, 7, 8, 9, 10] have evolved from CNNs and Transformers to generative VAEs [8] and diffusion models [9]. As VAEs often over-smooth details and diffusion models incur high latency, hybrid CNN-Transformer architectures [10]appears as a promising direction, as they combine CNNs' efficient local feature extraction with Transformers' ability to capture long-range dependencies. However, existing hybrid models rely on a relatively shallow fusion strategy—for instance, extracting features through parallel branches and performing a one-off concatenation or addition at the end.

We argue that this "one-shot" fusion fails to fully exploit the synergy of the two architectures. CSI exhibits a *local-global duality*, where small-scale fading (local) overlaid on large-scale shadowing (global) structures. Thus, it demands a deeper, more dynamic interaction mechanism. How to design an architecture that allows CNN's local insights to continuously guide Transformer's global modeling, while Transformer's global view, in turn, regularizes CNN's fine-detail generation, remains unresolved.

In this paper, we present the Bi-directional GAN for CSI Extrapolation Scheme (BiG-CES), which implements a principle of deep synergy to mitigate the inherent weaknesses of

This work was supported in part by the Guangdong ST Programme under Grant 2024B0101020004; in part by the National Natural Science Foundation of China under Grant 62472452; in part by Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515010262, 2023B1515120058); in part by RGC under AoE/E-601/22-R.

[†]Qianyi Huang is the corresponding author.

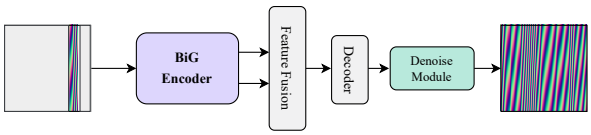

Fig. 1: The overall architecture of our framework. It consists of our core **BiG encoder**, a feature fusion stage, decoder and a denoise module for final refinement.

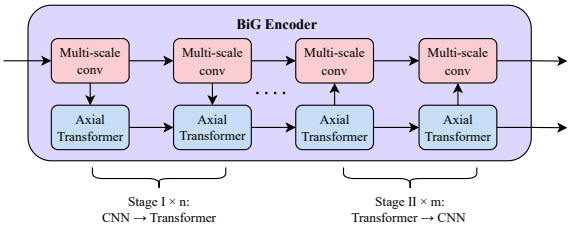

Fig. 2: The detailed architecture of **BiG Encoder**. The input consists of the narrowband CSI $\mathbf{X}_{\text{in}}$ and mask $\mathbf{M}$. It synergizes a parallel CNN branch (top) and a Transformer branch (bottom) through a two-stage, bidirectional cross-attention mechanism (vertical arrows).

each branch by leveraging the strengths of the other. On one hand, a pure Transformer, lacking fine-grained local details, can be hard to train on limited data, often failing to distinguish meaningful global patterns from noise. On the other hand, a pure CNN is constrained by its limited receptive field, failing to capture long-range dependencies. Our two-stage bidirectional cross-attention orchestrates a structured "dialogue" to resolve this. *In Stage I*, the CNN provides locally-structured features, forcing the Transformer to anchor its global attention on physically-grounded patterns. *In Stage II*, having established a reliable global context, the Transformer provides top-down guidance to regularize the CNN's detail generation, ensuring that local features align with the global structure. It is this structured, iterative refinement that empowers our model with the robustness to maintain high fidelity across wide frequency bands.

The main contributions of this work are: 1) a novel GAN-based BiG-CES architecture optimized for real-time wideband CSI extrapolation; 2) a two-stage bi-directional cross-attention mechanism for deep local-global feature synergy; and 3) a rigorous Sim-to-Real validation demonstrating up to 4dB SNR gain and a reduction of real-world UAV ranging error from 3.81m to 0.56m (a nearly seven-fold improvement).

2. PROBLEM FORMULATION

Channel Model and Sensing Resolution. In OFDM-based wireless systems, the frequency-domain CSI precisely describes the physical channel. For a multipath environment, the complex channel response $H(f_i)$ at the i-th subcarrier is given by: $H(f_i) = \sum_{l=1}^{L} \alpha_l e^{-j2\pi f_i \tau_l}$, where α_l and τ_l are the complex attenuation and delay of the l-th path, which are frequency-dependent in wideband scenarios [11]. A key performance metric, the range resolution Δd, is fundamentally limited by the signal bandwidth B, following $\Delta d \approx c/B$ [12]. This underscores the necessity of wideband signals for high-precision sensing applications.

Problem Formulation via CSI Imagization. The core task of this paper is to extrapolate a full wideband CSI matrix, $\mathbf{H}_{\text{full}} \in C^{N_s \times N_t}$, from a known narrowband CSI, $\mathbf{H}_{\text{known}}$, where N_s is the number of subcarriers and N_t is the number of time snapshots. We treat the complex-valued CSI as a two-channel spectral image, $\mathbf{X}_{\text{image}} \in R^{N_s \times N_t \times 2}$, by stacking its real and imaginary parts [13]. It is important to note that these CSI images possess unique physical properties, such as high-frequency oscillatory patterns and strong spatial correlations, which differ significantly from natural images [14]. This motivates the need for a bespoke architecture rather than off-the-shelf vision models.

Under this representation, our goal transforms into learning a generative mapping G [10] that can produce a complete and high-fidelity wideband CSI $\widehat{\mathbf{X}}_{\text{out}}$, conditioned on the input image of the known narrowband, $\mathbf{X}_{\text{in}}$ and a binary mask $\mathbf{M}$ ($\mathbf{M}$ is 1 for known subcarriers and 0 otherwise):

$$\widehat{\mathbf{X}}\text{out} = \mathbf{M} \odot \mathbf{X}_{\text{in}} + (\mathbf{1} - \mathbf{M}) \odot G(\mathbf{X}_{\text{in}}, \mathbf{M}), \qquad (1)$$

3. MODEL DESIGN

Our framework is engineered to address a fundamental challenge in CSI extrapolation: the dual and variable nature of the input data. CSI exhibits a *local-global duality*, where fine-grained local variations are superimposed on long-range global patterns. Furthermore, the manifestation of these properties varies dramatically with the physical scenario. A truly robust model must therefore adaptively handle both this structural duality and scenic variability. Our generator architecture is designed to tackle these two challenges.

3.1. Generator Architecture

The generator's design is a direct response to the intrinsic complexity of CSI data. BiG-CES, which follows an encoder-decoder paradigm (as shown in Fig. 1), orchestrates a synergistic process of parallel feature extraction and iterative fusion to handle local and global structures concurrently.

The Dual-Branch Encoder: At the heart of our generator lies a parallel dual-branch encoder (Fig. 2), where a CNN branch and a Transformer branch process the input features simultaneously. The *CNN Branch*, acting as an adaptive local feature expert, is built upon a cascade of *Multi-scale Convolutional Blocks*. Each block consists of three parallel 3×3 convolutional layers with dilation rates of 1, 2, and 4, respectively, each followed by a ReLU activation. The outputs

are then dynamically merged via a learned gating mechanism. This multi-scale design is crucial for handling scenic variability, as it allows the network to adaptively adjust its receptive field to capture the most salient local correlations. The *Transformer Branch*, serving as the global structure expert, is composed of a series of what we term *Axial-Attention Swin-Transformer Blocks* [15]. By factorizing the 2D self-attention into two sequential 1D operations along the spatial axes [16], it efficiently models long-range dependencies while remaining computationally tractable.

Bidirectional Cross-Attention: The key to unlocking the synergy between the two branches is *Bidirectional Cross-Attention* mechanism, an iterative, bidirectional fusion strategy that unfolds in two stages across the encoder's depth. In the initial *Stage I*, spanning n blocks, the two branches evolve in parallel, after which a cross-attention module is employed where the query (Q) comes from the Transformer branch, while the key (K) and value (V) come from the CNN branch. This process injects the stable, scene-adapted local features from the CNN into the Transformer, providing a robust anchor for its global modeling. In the subsequent *Stage II*, spanning m blocks, the roles are reversed. The query (Q) now comes from the CNN branch, with the key (K) and value (V) supplied by the Transformer branch. This injects the global context captured by the Transformer back into the CNN, ensuring that the generated local details adhere to a globally coherent structure. This two stage, bi-directional refinement is the essence of the BiG-CES design.

Feature Fusion: Upon exiting the BiG Encoder, the final feature maps from the CNN branch and the Transformer branch are synergized through the Feature Fusion module. Specifically, we employ a learnable gating mechanism that adaptively combines the two feature streams, allowing the model to dynamically weigh the importance of local versus global information.

Decoder with a Denoise Module. To improve reconstruction quality and robustness, our decoder incorporates a Denoise Module. The function of this module is to treat a range of non-ideal factors—such as reconstruction artifacts, sim-to-real domain gaps, and measurement noise—as a single, generalized "noise" term that can be filtered out [17]. It allows the generator to focus on synthesizing the primary channel structure while this module handles the filtering of various error sources. Consequently, this design improves the model's ability to generalize across different scenarios and SNR conditions.

3.2. Discriminator Architecture

Our discriminator is a standard PatchGAN architecture [18]. with four 4×4, stride-2 convolutional layers. To ensure training stability, each layer is followed by Spectral Normalization [19] and a LeakyReLU activation with a negative slope of 0.2.

3.3. Training Objective

Our model is trained via the adversarial objective of WGAN-GP [20], which formulates a min-max game between the generator (G) and the discriminator (D). The discriminator's objective, $\mathcal{L}_D$, aims to distinguish real from generated samples while being regularized by a gradient penalty. Concurrently, the generator's optimization is guided by a composite loss function, $\mathcal{L}_G$, designed to balance adversarial feedback with reconstruction accuracy:

$$\mathcal{L}_G = \lambda_1 \mathcal{L}_1 + \lambda_{\text{p}} \mathcal{L}_{\text{perceptual}} + \lambda_{\text{s}} \mathcal{L}_{\text{style}} - \lambda_{\text{adv}} E[D(\widehat{\mathbf{X}}_{\text{out}})] \quad (2)$$

where the first three terms are reconstruction losses that ensure content fidelity and perceptual quality [21], while the final adversarial term rewards the generator for fooling the discriminator. The λ terms are weighting factors.

4. EXPERIMENTAL EVALUATION

This section first details the experimental setup, including the datasets, baselines, and evaluation metrics. We then present a quantitative comparison of CSI reconstruction performance, followed by an application-driven validation on a real-world UAV ranging task and ablation study.

4.1. Experimental Setup

Datasets and Signal Parameters. Our evaluation is conducted on two distinct datasets: a large-scale simulated dataset and a real-world UAV dataset [22]:

1) Simulated Dataset: Following the methodology in HORCRUX [5], we generated 200,000 channel samples based on the geometric multipath model. The center frequency is 3.75 GHz and the total bandwidth is 80 MHz, comprising 1280 active subcarriers. The simulation covered four scenarios with increasing multipath richness as defined by 3GPP TR 38.901: outdoor-open (2-5 paths), urban (5-15 paths), indoor (10-30 paths), and dense-urban (>30 paths) [11]. The data was split into 196,000 for training, 2,000 for validation, and 2,000 for testing. For our extrapolation task, the full 80 MHz CSI served as the ground truth, while a 20 MHz subset (320 subcarriers) was used as the narrowband input.

2) Real-World UAV Dataset: To assess performance in a practical ISAC scenario, we utilized a dataset captured in a 16 km^2 urban environment [22]. The setup consisted of a UAV-mounted transmitter and four stationary ground receivers. Centimeter-level ground truth was provided by a RTK-GNSS system. This dataset served exclusively as the test set, where the model is trained on the simulated dataset and directly tested on this real-world dataset. For the downstream ranging application, we compared the ranging performance using the original 20 MHz observed CSI against the 80 MHz extrapolated CSI (extrapolated from the same 20MHz CSI by BiG-CES) across four scenarios defined by

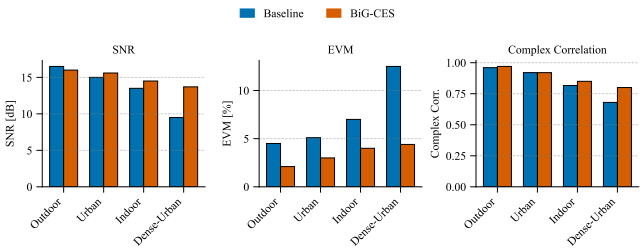

Fig. 3: Performance comparison of BiG-CES against the baseline across four simulated scenarios.

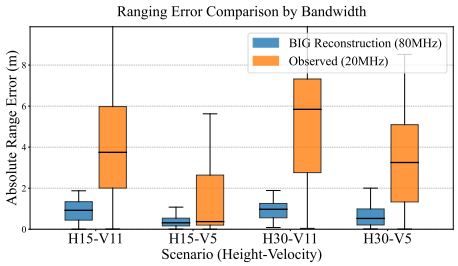

Fig. 4: Ranging performance comparison on the real-world dataset. The box plot shows the distribution of absolute range errors for the 20 MHz CSI (observed) versus the 80 MHz CSI (extrapolated) by BiG-CES across four flight scenarios.

combinations of maximum altitude (15 m and 30 m) and velocity (5 m/s and 11 m/s).

Baselines and Implementation Details. Our primary baseline is the state-of-the-art HORCRUX model [5], and we evaluate performance using EVM, SNR, and Complex Correlation (ρ_C). The BiG-CES framework was implemented in PyTorch and trained for 1.2 million iterations on a single NVIDIA A100 GPU. We employed the ADAM optimizer with a batch size of 2 and learning rates of 5×10^{-5} for the generator and 1×10^{-4} for the discriminator, which was updated half as frequently as the generator.

4.2. Performance Evaluation

CSI Reconstruction Performance: As shown in Fig. 3, BiG-CES consistently outperforms HORCRUX, with the performance gap widening significantly as scenario complexity increases. As the channel environment transitions from sparse to structurally complex, the baseline struggles to capture fine-grained channel features, leading to a noticeable performance degradation. In stark contrast, our model demonstrates excellent robustness, as its SNR and correlation metrics show only a minor decay across all scenarios. This stable performance maximizes its advantage in the most challenging dense-urban scenario, achieving a **4.5** dB SNR improvement and reducing the EVM from over 12% to around **3.5%**.

UAV Ranging on Real-world Dataset: To validate the utility of the extrapolated CSI, we compared the performance of a complete ranging pipeline when using the 20 MHz CSI (observed) versus our extrapolated 80 MHz CSI. The pipeline first enhances the target SNR on the Range-Doppler (RD) map via static clutter suppression and a long coherent processing interval (CPI), followed by a 2D Cell-Averaging Constant False Alarm Rate (CA-CFAR) detector for target detection, and finally a Kalman filter for robust tracking. The results in Fig 4 show that using the 80 MHz extrapolated CSI from BiG-CES reduces the mean absolute error (MAE) to just 0.56 meters, a nearly seven-fold improvement over the 3.81-meter MAE from the observed 20 MHz CSI. The box plot confirms this substantial gain, showing dramatic reductions in both median error and variance across all four scenarios. This result validates that our synthesized bandwidth can significantly improve sensing performance.

Table 1: Ablation study on the Dense-Urban dataset.

Model Configuration	SNR (dB) ↑	EVM (%) ↓	Complex Corr ↑
Full Model (Proposed)	**13.75**	**4.21**	**0.80**
w/o Cross-Attention	11.98	5.82	0.73
CNN Branch Only	10.15	8.54	0.61
Transformer Branch Only	10.87	7.66	0.66
w/o Denoise Module	12.35	5.58	0.75

Ablation Study. To dissect each component's contribution to BiG-CES, we conducted the ablation study on the challenging dense-urban scenario (summarized in Table 1). Replacing the bidirectional cross-attention mechanism ('w/o Cross-Attention') with a simple feature addition results in a 1.77 dB SNR drop, confirming the necessity of our iterative fusion strategy. As expected, using only the CNN or Transformer branch leads to a catastrophic performance collapse, validating the dual-branch design. Furthermore, removing the Denoise Module incurs a significant performance penalty of 1.4 dB in SNR, highlighting its critical role in enhancing the final reconstruction quality and robustness. These results validate that each component is not only effective but indispensable.

5. CONCLUSION

We propose the BiG-CES framework and validate that it successfully extrapolates high-fidelity wideband CSI from narrowband observations. BiG-CES centers on a dual-branch generator and a two-stage bi-directional cross-attention mechanism that adaptively handles the local-global duality of CSI data across diverse physical scenarios. Experimental results not only show significant gains over SOTA methods in numerical metrics, but also demonstrate substantial real-world value by enabling an autonomous UAV ranging system to achieve a 0.56-meter mean absolute error, compared to the 3.81m error with narrowband CSI. This research provides a viable, data-driven solution for unlocking high-resolution ISAC sensing capabilities under the spectral constraints of communication system, paving the way for 6G evolution.

6. REFERENCES

[1] Fan Liu, Yuanhao Cui, Christos Masouros, Jie Xu, Tony Xiao Han, Yonina C Eldar, and Stefano Buzzi, "Integrated sensing and communications: Toward dual-functional wireless networks for 6g and beyond," *IEEE journal on selected areas in communications*, vol. 40, no. 6, pp. 1728–1767, 2022.

[2] Mohammed Khaled Banafaa, Ömer Pepeoğlu, Ibraheem Shayea, Abdulraqeb Alhammadi, Zaid Ahmed Shamsan, Muneef A Razaz, Majid Alsagabi, and Sulaiman Al-Sowayan, "A comprehensive survey on 5g-and-beyond networks with uavs: Applications, emerging technologies, regulatory aspects, research trends and challenges," *IEEE access*, vol. 12, pp. 7786–7826, 2024.

[3] Chao-Kai Wen, Wan-Ting Shih, and Shi Jin, "Deep Learning for Massive MIMO CSI Feedback," in *2018 IEEE Globecom Workshops (GC Wkshps)*, 2018.

[4] Deepak Vasisht, Swarun Kumar, Hariharan Rahul, and Dina Katabi, "Eliminating channel feedback in next-generation cellular networks," in *Proceedings of the 2016 ACM SIGCOMM Conference*, 2016, pp. 398–411.

[5] Avishek Banerjee, Xingya Zhao, Vishnu Chhabra, Kannan Srinivasan, and Srinivasan Parthasarathy, "Horcrux: Accurate cross band channel prediction," in *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, 2024, pp. 1–15.

[6] Matthias Pätzold, *Mobile radio channels*, John Wiley & Sons, 2011.

[7] Yang Xu, Mingqi Yuan, and Man-On Pun, "Transformer empowered csi feedback for massive mimo systems," in *2021 30th Wireless and Optical Communications Conference (WOCC)*. IEEE, 2021, pp. 157–161.

[8] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[10] Xiaojun Bi, Shuo Li, Changdong Yu, and Yu Zhang, "A novel approach using convolutional transformer for massive mimo csi feedback," *IEEE Wireless Communications Letters*, vol. 11, no. 5, pp. 1017–1021, 2022.

[11] NTT Docomo et al., "5g channel model for bands up to100 ghz," Tech. Rep., Technical report, 2016.

[12] Andrea Goldsmith, *Wireless communications*, Cambridge university press, 2005.

[13] Tianqi Wang, Chao-Kai Wen, Shi Jin, and Geoffrey Ye Li, "Deep learning-based csi feedback approach for time-varying massive mimo channels," *IEEE Wireless Communications Letters*, vol. 8, no. 2, pp. 416–419, 2018.

[14] Jiajia Guo, Chao-Kai Wen, Shi Jin, and Geoffrey Ye Li, "Overview of deep learning-based csi feedback in massive mimo systems," *IEEE Transactions on Communications*, vol. 70, no. 12, pp. 8017–8045, 2022.

[15] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012–10022.

[16] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen, "Axial-deeplab: Stand-alone axial-attention for panoptic segmentation," in *European conference on computer vision*. Springer, 2020, pp. 108–126.

[17] Singanallur V Venkatakrishnan, Charles A Bouman, and Brendt Wohlberg, "Plug-and-play priors for model based reconstruction," in *2013 IEEE global conference on signal and information processing*. IEEE, 2013, pp. 945–948.

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[19] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida, "Spectral normalization for generative adversarial networks," in *International Conference on Learning Representations*, 2018.

[20] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville, "Improved training of wasserstein gans," *Advances in neural information processing systems*, vol. 30, 2017.

[21] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[22] Julia Beuster, Carsten Andrich, Michael Döbereiner, Steffen Schieler, Maximilian Engelhardt, Christian Schneider, and Reiner Thomä, "Measurement testbed for radar and emitter localization of uav at 3.75 ghz," in *2023 17th European Conference on Antennas and Propagation (EuCAP)*. IEEE, 2023, pp. 1–5.