Robust Channel Representation for Wireless: A Multi-Task Masked Contrastive Approach

Berkay Guler[1] **Giovanni Geraci**[2] **Hamid Jafarkhani**[1]
[1]Center for Pervasive Communications and Computing, University of California, Irvine
[2]Nokia Standards & Universitat Pompeu Fabra, Barcelona, Spain

Abstract

Wireless environments present unique challenges for machine learning (ML) due to partial observability, multi-domain characteristics, and dynamic nature. To address these challenges, we propose ContraWiMAE (Wireless Masked Contrastive Autoencoder), a multi-task learning framework that learns robust representations from incomplete wireless channel observations avoiding expensive augmentation engineering. Our approach combines masked autoencoder learning with a novel masked contrastive objective that contrasts differently masked versions of the same channel, leveraging the inherent wireless complexity as natural augmentation. We employ a curriculum learning strategy systematically developing representations that maintain structural properties while enhancing discriminative capabilities. The framework enables learning task-specific invariances, which we demonstrate through noise invariance and improved linear separability tested on channel estimation and cross-frequency beam selection in unseen environments. Our experiments demonstrate superior representation stability under severe noise conditions with performance close to supervised training.

1 Introduction

Contemporary ML research, dominated by vision and language, relies on clean training data and tests robustness through artificially generated corruptions and adversarial attacks. Wireless communications presents a contrast where corruption is fundamental to the domain, due to practical limitations like sparse pilot placement and physical realities such as noise and interference. While these inherent imperfections create challenging conditions for conventional ML algorithms, they simultaneously establish wireless as an exceptional testbed for developing truly robust representation learning techniques. We argue that this intrinsic complexity should drive the design of next-generation wireless ML methods.

While self-supervised learning has achieved remarkable success through masked prediction [1–3] and contrastive learning [4–7], recent wireless adaptations [8–19] exhibit complementary limitations. Contrastive methods require careful augmentation engineering and positive-negative pair selection, involve multiple forward passes, and often capture discriminative features while missing crucial structural channel characteristics. Reconstructive approaches adopt language and vision architectures with minimal wireless-specific consideration, often emphasizing low-level statistical patterns over the semantic channel features required for downstream tasks.

Building on our robustness-by-design philosophy, we propose leveraging the inherent complexity of wireless channels through a multi-task framework that combines masked autoencoders with a novel masked contrastive objective. Random masking serves as a natural augmentation for wireless channels, where partial observability is inherent due to pilot limitations.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI and ML for Next-Generation Wireless Communications and Networking (AI4NextG).

Aggressive masking drives the model to capture small-scale fading patterns while maintaining structural coherence without manual augmentation design. We complement this with a novel masked contrastive objective that learns invariant representations by contrasting differently masked versions of original and augmented channel samples, demonstrating its effectiveness through noise invariance.

Unlike existing contrastive methods that contrast different views of a sample, we contrast differently masked versions of the same channel, enabling the model to recognize consistent propagation patterns from different partial observations. This approach naturally captures both structural coherence and discriminative features.

Our contributions include the following:

- **Masked contrastive learning**: Our masked contrastive approach learns invariant representations by comparing different partial observations of the same underlying channel.
- **Two-stage curriculum learning**: Our two-stage training approach builds contrastive learning on pretrained masked autoencoder foundations, maintaining structural coherence while improving discriminative power.
- **Natural augmentation through masking**: Our masking approach leverages the inherent complexity of wireless channels (noise, fading, partial observability) as augmentation and eliminates the need for expensive augmentation engineering.

Experiments on ray-traced wireless channels demonstrate that ContraWiMAE[1] (Wireless Masked Contrastive Autoencoder) learns robust, noise-invariant representations with performance rivaling supervised training across channel estimation and beam selection tasks.

2 ContraWiMAE Framework

We consider a complex-valued channel matrix $\mathbf{H} \in \mathbb{C}^{N_\mathrm{s} \times N_\mathrm{f}}$ representing the frequency response across N_s spatial dimensions and N_f frequency dimensions. We formulate our channel representation problem as learning an encoder $f_\theta : \mathbb{C}^{N_\mathrm{s} \times N_\mathrm{f}} \to \mathbb{R}^d$ that produces d-dimensional vectors robust to: (1) missing observations due to pilot patterns, (2) noise variations across different SNR regimes, and (3) distribution shifts between training and deployment environments.

Architecture Design. We employ an asymmetric encoder-decoder architecture based on masked autoencoders [2] with adaptations for complex-valued channel inputs. The channel matrix $\mathbf{H} \in \mathbb{C}^{N_\mathrm{s} \times N_\mathrm{f}}$ is partitioned into non-overlapping patches of size $(p_\mathrm{s} \times p_\mathrm{f})$, resulting in P patches. Each channel patch is flattened row-wise into a column vector $\mathbf{h}_i \in \mathbb{C}^{p_\mathrm{s} \cdot p_\mathrm{f}}$, creating the patch vector $\mathbf{h} = [\mathbf{h}_1; \mathbf{h}_2; \ldots; \mathbf{h}_P] \in \mathbb{C}^{P \cdot p_\mathrm{s} \cdot p_\mathrm{f}}$. Given masking ratio r_m, we define $\mathbf{m} \in \{0,1\}^P$ as a random binary vector where exactly $\lfloor r_\mathrm{m} \cdot P \rfloor$ entries are 1. The full element-wise mask is $\tilde{\mathbf{m}} = \mathbf{m} \otimes \mathbf{1}_{p_\mathrm{s} \cdot p_\mathrm{f}} \in \{0,1\}^{P \cdot p_\mathrm{s} \cdot p_\mathrm{f}}$ where $\otimes$ denotes the Kronecker product. The visible elements are extracted to form $\mathbf{h}_\mathrm{vis} \in \mathbb{C}^{N_\mathrm{vis}}$ by selecting unmasked entries from $\mathbf{h}$, which are converted to real-valued representation $\mathbf{h}_\mathrm{real} = [\Re(\mathbf{h}_\mathrm{vis}); \Im(\mathbf{h}_\mathrm{vis})] \in \mathbb{R}^{2N_\mathrm{vis}}$ where $N_\mathrm{vis} = \|\mathbf{1} - \tilde{\mathbf{m}}\|_1$. The encoder f_θ processes $\mathbf{h}_\mathrm{real}$, while a lightweight decoder g_ϕ reconstructs the channel from encoder output and mask information $\mathbf{m}$. The asymmetry forces rich encoder representations since the lightweight decoder cannot compensate for poor feature extraction. To achieve this, we propose a multi-task objective $\mathcal{L} = \alpha \mathcal{L}_\mathrm{r} + (1-\alpha)\mathcal{L}_\mathrm{c}$ where α balances structural learning with discriminative learning and $\mathcal{L}_\mathrm{r}$ and $\mathcal{L}_\mathrm{c}$ are defined below.

Reconstruction Objective. The decoder reconstructs all patches but is optimized for masked patches:

$$\mathcal{L}_\mathrm{r} = \mathbb{E}_{\mathbf{h},\mathbf{m}} \left[\|\tilde{\mathbf{m}} \odot \mathbf{h} - \tilde{\mathbf{m}} \odot \hat{\mathbf{h}}\|_2^2 \right],$$

where $\hat{\mathbf{h}} = g_\phi(f_\theta(\mathbf{h}_\mathrm{real}), \mathbf{m})$ is the reconstructed channel vector and $\odot$ is the element-wise product.

Masked Contrastive Objective. For each channel $\mathbf{H}_i$, we generate a positive pair by adding noise $\mathbf{H}_i^+ = \mathbf{H}_i + \mathbf{N}$ where $\mathbf{N} \in \mathbb{C}^{N_\mathrm{s} \times N_\mathrm{f}}$ is a matrix of i.i.d. complex Gaussian random variables with $N_{k,\ell} \sim \mathcal{CN}(0, \sigma_i^2)$ and applying different random masks $\mathbf{m}_i$ and $\mathbf{m}_i^+$. Encoded representations are projected to a contrastive space with $h_\psi : \mathbb{R}^d \to \mathbb{R}^{d_\mathrm{c}}$ producing $\mathbf{u}_i = h_\psi(f_\theta(\mathbf{h}_{i,\mathrm{real}}))$ and

[1]We make ContraWiMAE along with our dataset publicly available at: https://github.com/BerkIGuler/WirelessContrastiveMaskedLearning

$\mathbf{u}_i^+ = h_\psi(f_\theta(\mathbf{h}_{i,\text{real}}^+))$. We concatenate representations from both views across the batch to form $\mathbf{Z} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_B, \mathbf{u}_1^+, \mathbf{u}_2^+, \ldots, \mathbf{u}_B^+] \in \mathbb{R}^{2B \times d_\text{c}}$. Following the InfoNCE framework [4, 20], we treat each sample and its augmented version as positive pairs while all other samples serve as negatives:

$$\mathcal{L}_\text{c} = -\frac{1}{2B} \sum_{k=1}^{2B} \log \frac{\exp(\text{sim}(\mathbf{z}_k, \mathbf{z}_{\text{pos}(k)})/\tau)}{\sum_{j=1, j\neq k}^{2B} \exp(\text{sim}(\mathbf{z}_k, \mathbf{z}_j)/\tau)},$$

where $\mathbf{z}_k$ denotes the k-th row of $\mathbf{Z}$, $\text{pos}(k) = k + B$ if $k \leq B$, otherwise $\text{pos}(k) = k - B$, and $\text{sim}(\cdot, \cdot)$ denotes cosine similarity.

By contrasting differently masked versions of the same channel, this objective learns representations that are invariant to particular antennas and subcarriers while remaining noise invariant across varying corruption levels. This approach prevents overfitting to both noise artifacts in training data and specific measurement patterns or antenna configurations, while achieving data efficiency by generating multiple views from each channel sample. The resulting representations maintain semantic consistency, ensuring partial observations of the same channel map to similar feature spaces, and provide task agnostic learning that transfers across downstream applications like channel estimation and beamforming. Since masking and noise corruption naturally occur in wireless systems through varying pilot patterns and measurement uncertainty, the learned invariances directly enhance robustness in real deployment scenarios. The practical impact is improved performance under incomplete channel state information, making the framework suitable for diverse wireless environments where perfect channel observations are rarely available.

3 Experiments

Dataset. We use the DeepMIMO [21, 22] ray-tracing dataset to generate 2.5 million channels at 3.5 GHz across 56 diverse urban scenarios for pretraining. Additionally, we create a task dataset with 0.55 million channels at the same frequency across 10 new urban scenarios (see appendix A for details). For six of these scenarios, we also generate channels at 28 GHz. Each scenario features multiple single-antenna user equipments (UEs) uniformly placed on a grid, 1 meter apart with fixed height. Each base station (BS) is equipped with a uniform linear array containing $N_\text{s} = 32$ antenna elements spaced $\lambda/2$ apart. For scenarios with multiple BSs, we generate individual channels for each UE-BS pair. We employ 32 subcarriers with 30 kHz subcarrier spacing.

Training. We implement f_θ and g_ϕ as standard transformer models [23, 24] with learnable positional encodings. The projection head h_ψ is implemented as a two-layer MLP with $2d_\text{c}$ hidden dimension size and ReLU activation, preceded by average pooling over individual patch embeddings. We configure the encoder with 12 layers and 16 self-attention heads, while adopting a more compact decoder with only 4 layers and 8 self-attention heads, resulting in a compact architecture with 0.4M parameters due to the moderate input channel size. Both the embedding dimension and contrastive space dimensions are set to 64. We use a patch size of $(p_\text{s} \times p_\text{f}) = (16 \times 1)$ across spatial and frequency dimensions, yielding an input sequence length of 128. The masking ratio is set to $r_\text{m} = 0.9$ after experimenting with different masking ratios, with $\alpha = 0.9$ and $\tau = 0.2$. We train ContraWiMAE for 3000 epochs with a batch size of 8096 using the AdamW [25] optimizer. The learning rate schedule begins with a warm-start at 3e-6, linearly increasing to 3e-4 over the first 10 epochs, followed by cosine decay to 3e-6 by epoch 3000. For each batch, we uniformly sample an SNR value from the range $[5, 40]$ dB and select the corresponding σ_i^2 value accordingly.

To highlight the structural and discriminative properties of the learned representations, we evaluate ContraWiMAE's performance on two fundamentally different tasks: beam selection and channel estimation. These experiments evaluate complementary aspects of ContraWiMAE. Together, they validate the model's ability to learn both transferable and robust representations for diverse wireless tasks.

Cross-frequency Beam Selection. Channel state information (CSI) acquisition at mmWave frequencies is computationally expensive and time-consuming due to extensive beam sweeping requirements. Leveraging CSI from sub-6 GHz frequencies for mmWave beamforming is therefore highly desirable in modern multi-band communication systems. To evaluate the generalizability of ContraWiMAE representations across frequencies, we test whether a model pretrained on 3.5 GHz channels can effectively perform beam selection at 28 GHz without any access to 28 GHz labels. We employ a

k-nearest neighbor (kNN) classifier to provide a parameter-free evaluation of representation quality that directly measures embedding space structure without introducing additional optimization bias, which is particularly important for evaluating cross-frequency and cross-scenario generalization. Our approach works as follows: we first train a kNN classifier using the frozen ContraWiMAE encoder as a feature extractor on a split of the task dataset from eight scenarios containing 3.5 GHz channels paired with their corresponding beam labels, while reserving two other scenarios for testing. We then evaluate this trained classifier on 3.5 GHz channels with 28 GHz beam labels as ground truth. For label generation, we employ codebooks of sizes 16, 32, and 64 with uniform beam patterns and select the beam that maximizes the received power as the optimal choice.

We compare ContraWiMAE with a supervised baseline that uses the same encoder architecture but trains directly on 3.5 GHz channel-beam label pairs for each codebook size separately. The supervised baseline employs a randomly initialized encoder with a linear classifier head, trained end-to-end on the pretraining data. After training, we remove the classifier head and use the encoder as a feature extractor for kNN evaluation, keeping consistency with our evaluation protocol. We also assess noise robustness by evaluating both ContraWiMAE and the supervised baseline on 3.5 GHz channels corrupted with additive white Gaussian noise at SNR levels from 0 to 40 dB in 5 dB increments. To evaluate the contribution of our masked contrastive objective, we include a reconstruction-only baseline that trains using only the masked autoencoder loss ($\alpha = 1$) and follows the same evaluation protocol. For completeness, we also report the results of the same-frequency beam selection in the next section where both pairs of channel-beam labels are obtained at 3.5 GHz for the evaluation phase.

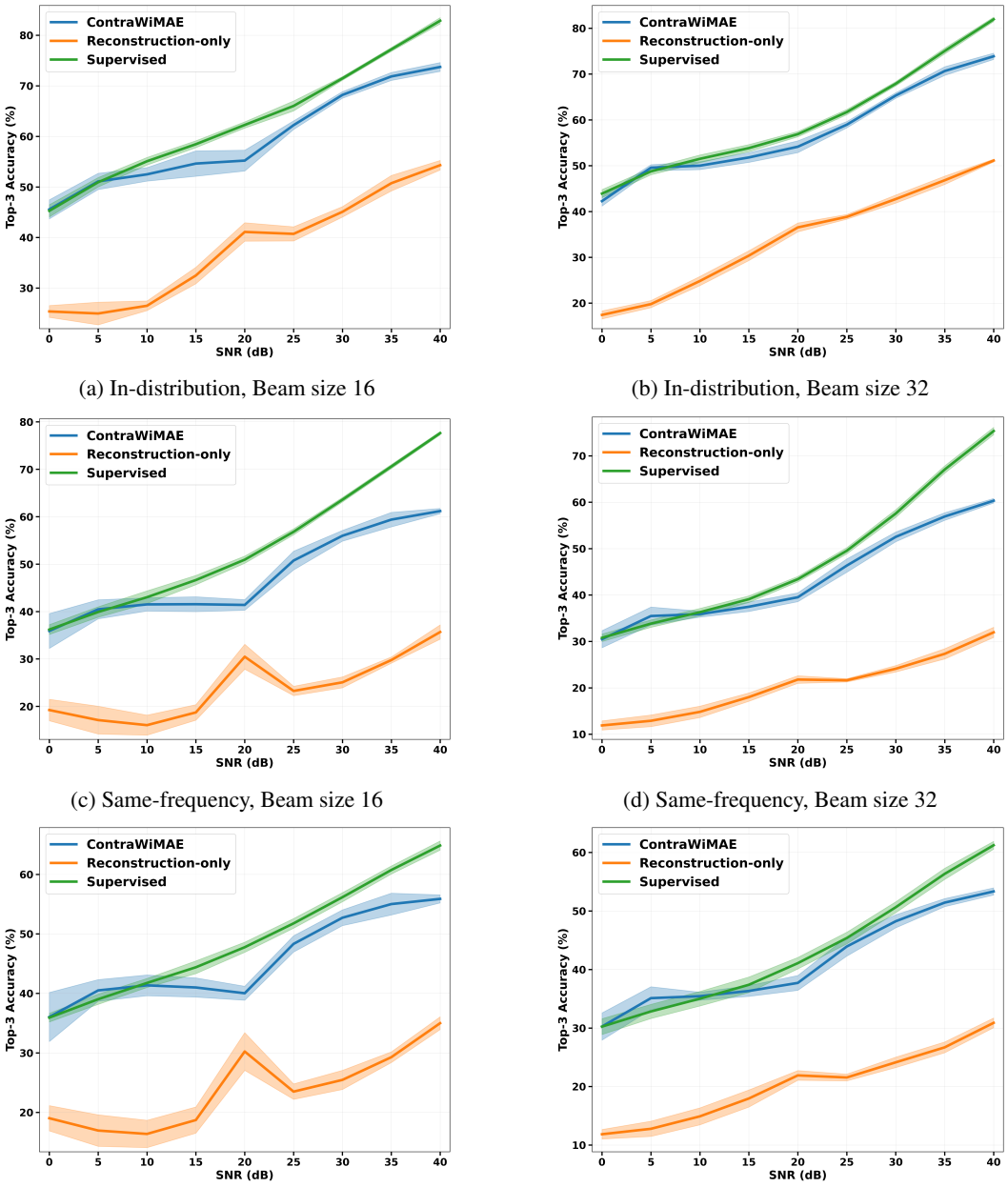
Channel Estimation. We consider an uplink channel estimation scenario where the UE transmits a pilot symbol only on a single subcarrier. The BS estimates the channel at each antenna element with measurement errors modeled as Gaussian noise. The BS then employs the pretrained ContraWiMAE encoder-decoder pair to reconstruct the entire channel from this noisy partial observation. We report the normalized mean square error (NMSE) on the same unseen test scenarios from the task dataset (identical to the beam selection task) across varying test SNR levels from 0 to 40 dB in 5 dB increments.

We compare ContraWiMAE against several baselines: (1) a supervised encoder-decoder model trained on the pretraining dataset to minimize reconstruction error for the fixed pilot pattern, (2) a supervised model trained on noisy data from the pretraining dataset at the same SNRs as the test set, and (3) a linear interpolation baseline. We also finetune ContraWiMAE on the fixed pilot pattern using the remaining scenarios from the task dataset, experimenting with finetuning budgets of 2.5%, 10%, 25%, 50%, 75%, and 100%, where 100% corresponds to finetuning on all remaining task scenarios.

4 Results and Analysis

We consider different forms of distribution shifts to evaluate the robustness of ContraWiMAE representations. Figures 1a and 1b demonstrate the noise robustness of our representations by fixing the environment and frequency while observing noise effects. To evaluate sensitivity to sample selection, we repeatedly select random subsets of samples from the training and test scenarios and report the average top-3 accuracy with $k = 10$ across different codebook sizes. Our in-distribution evaluation shows that ContraWiMAE consistently closes the performance gap with the supervised model as noise conditions become more challenging, even matching its performance in the 0-10 dB range.

Figures 1c and 1d show out-of-distribution (OOD) performance evaluated on test scenarios non-overlapping with kNN training thus isolating environment shift effects. ContraWiMAE maintains its competitive edge over the supervised model at challenging SNRs, while all models experience similar performance deterioration. As shown in Figures 1e and 1f, we evaluate cross-frequency representation robustness by predicting beam labels on unseen test scenarios at an unseen frequency simultaneously. Despite being a self-supervised model, because of its robustness to noise, our approach consistently outperforms the supervised baseline when environmental shifts and challenging noise conditions are coupled with cross-frequency test conditions. The reconstruction-only baseline performs poorly in the absence of our masked contrastive objective, establishing its effectiveness.



(a) In-distribution, Beam size 16

(b) In-distribution, Beam size 32

(c) Same-frequency, Beam size 16

(d) Same-frequency, Beam size 32

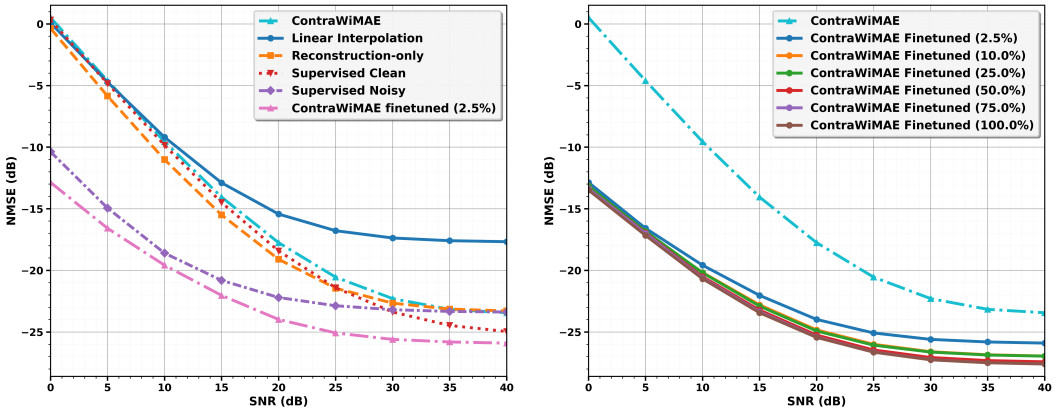
(e) Cross-frequency, Beam size 16

(f) Cross-frequency, Beam size 32

Figure 1: Beam selection performance under varying SNR for in-distribution (first row), OOD (second row), and severe OOD (third row) settings with beam sizes 16 and 32

We report channel estimation performance on the same unseen test scenarios in Figure 2a. Although ContraWiMAE and the reconstruction-only baseline are pretrained on a different set of scenarios with random masking, their performance is competitive with the supervised model trained on clean data using a single reconstruction mask matching the pilot position. We also observe that the dual learning objective of ContraWiMAE leads to only a minor sacrifice in reconstruction performance, as shown by its performance nearly matching the reconstruction-only baseline. As expected, the supervised model trained on the same noise levels as the test SNRs outperforms all models due to its learned noise behavior.

We further study the pilot-specific data scaling behavior of ContraWiMAE in Figure 2b, demonstrating that ContraWiMAE performance experiences diminishing gains with increasing finetune



(a) ContraWiMAE vs. baselines (b) ContraWiMAE finetuning sample complexity

Figure 2: Channel estimation on unseen scenarios

budget, where the finetuning gain becomes negligible above a 10% finetune data budget. Pretrained ContraWiMAE, when finetuned on 2.5% of the task dataset with only 7000 channels (three orders of magnitude smaller than the pretraining data), outperforms the supervised noise baseline at all SNR levels, demonstrating ContraWiMAE's superior sample efficiency. We note that our masked contrastive objective does not aim to improve reconstruction quality by design, but rather reorganizes the embedding space with better discriminative features while almost perfectly retaining reconstructive capability.

In Figure 3, we also include t-SNE (t-distributed Stochastic Neighbor Embedding) representations of ContraWiMAE and the reconstruction-only baseline across 0, 5, 10, and 15 dB SNR to visually demonstrate the enhanced noise robustness provided by the masked contrastive objective. Specifically, we color the channels based on the presence of a line-of-sight (LoS) path. Despite proper clustering behavior at 10 and 15 dB SNR by both models, ContraWiMAE retains clearer boundaries at 0 and 5 dB SNRs, where the reconstruction-only baseline heavily mistakes non-LoS channels for LoS.

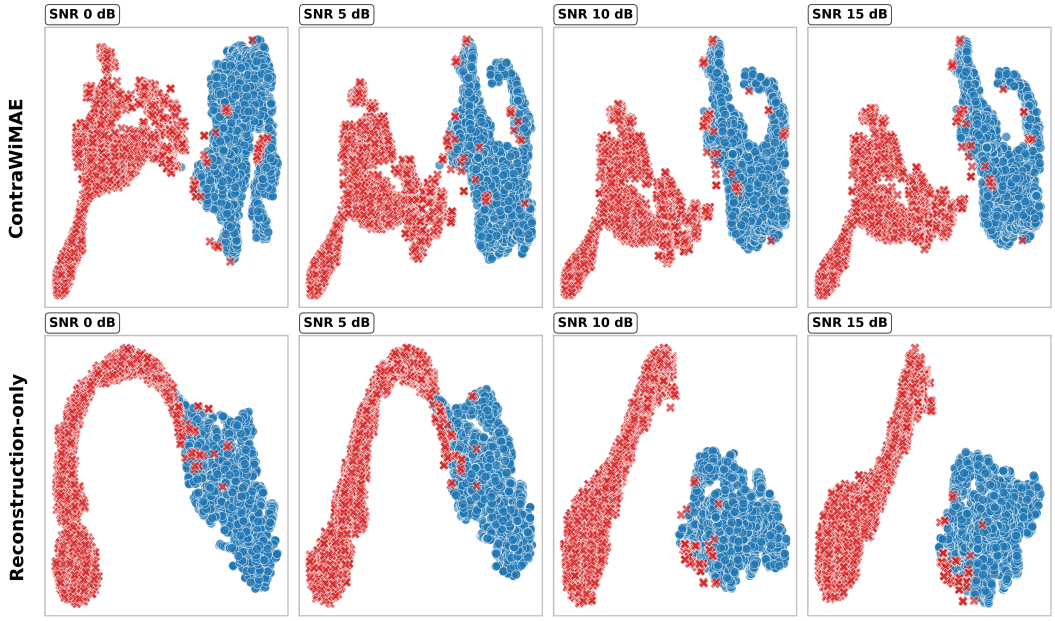


Figure 3: t-SNE representation with LoS (blue) and non-LoS (red) labels under varying SNR

5 Conclusion

We propose a multi-task framework that combines masked autoencoders with masked contrastive learning for learning robust wireless channel representations. By contrasting differently masked versions of the same channel with additive noise, our approach leverages wireless complexity as natural augmentation while learning both structural and discriminative features. Experiments on channel estimation and beam selection demonstrate performance comparable to supervised training with superior robustness under noise and distribution shifts. The framework achieves strong out-of-distribution results without or with minimal finetuning, highlighting its practical value for wireless applications where site-specific data is limited. Future directions include pretraining on datasets with diverse frequencies, system parameters, and antenna configurations to further validate generalization capabilities across a wide range of relevant downstream tasks.

Acknowledgments

B. Guler and H. Jafarkhani are with the Center for Pervasive Communications and Computing, University of California, Irvine CA, USA. They were supported in part by the NSF Award CNS-2229467.

G. Geraci is with Nokia Standards and Universitat Pompeu Fabra, Spain. He was supported in part by grants PID2021-123999OB-I00, PID2024-156488OB-I00, CEX2021-001195-M, and CNS2023-145384.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019, pp. 4171–4186.

[2] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *CVPR*, 2022, pp. 16 000–16 009.

[3] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," in *ICLR*, 2022.

[4] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020, pp. 1597–1607.

[5] J.-B. Grill, F. Strub *et al.*, "Bootstrap your own latent - a new approach to self-supervised learning," in *NeurIPS*, 2020, pp. 21 271–21 284.

[6] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020, pp. 9726–9735.

[7] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *NeurIPS*, 2020, pp. 9912–9924.

[8] P. Stephan, F. Euchner, and S. ten Brink, "Angle-delay profile-based and timestamp-aided dissimilarity metrics for channel charting," *IEEE Trans. Commun.*, vol. 72, no. 9, pp. 5611–5625, 2024.

[9] L. Yu, L. Shi, J. Zhang, J. Wang, Z. Zhang, Y. Zhang, and G. Liu, "ChannelGPT: A large model to generate digital twin channel for 6G environment intelligence," 2024.

[10] T. Yang, P. Zhang, M. Zheng, Y. Shi, L. Jing, J. Huang, and N. Li, "WirelessGPT: A generative pre-trained multi-task learning framework for wireless communication," 2025.

[11] V. Palhares, S. Taner, and C. Studer, "CSI2Vec: Towards a universal CSI feature representation for positioning and channel charting," 2025.

[12] J. Jiang, W. Yu, Y. Li, Y. Gao, and S. Xu, "A MIMO wireless channel foundation model via CIR-CSI consistency," 2025.

[13] Z. Liu, J. Chen, Y. Xu, T. Ma, J. Liu, H. Zhou, and D. Niyato, "Leveraging self-supervised learning for MIMO-OFDM channel representation and generation," 2024.

[14] T. Jiao, C. Ye, Y. Huang *et al.*, "6G-oriented CSI-based multi-modal pre-training and downstream task adaptation paradigm," in *ICC Workshops*, 2024, pp. 1389–1394.

[15] A. Zayat, M. A. Hasabelnaby, M. Obeed, and A. Chaaban, "Transformer masked autoencoders for next-generation wireless communications: Architecture and opportunities," *IEEE Commun. Mag.*, vol. 62, no. 7, pp. 88–94, 2024.

[16] G. Pan, K. Huang, H. Chen, S. Zhang, C. Häger, and H. Wymeersch, "Large wireless localization model (LWLM): A foundation model for positioning in 6G networks," 2025.

[17] B. Liu, S. Gao, X. Liu, X. Cheng, and L. Yang, "WiFo: Wireless foundation model for channel prediction," 2025.

[18] A. Salihu, M. Rupp, and S. Schwarz, "Self-supervised and invariant representations for wireless localization," *IEEE Trans. Wireless Commun.*, vol. 23, no. 8, pp. 8281–8296, 2024.

[19] S. Alikhani, G. Charan, and A. Alkhateeb, "Large wireless model (LWM): A foundation model for wireless channels," 2024.

[20] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2019.

[21] A. Alkhateeb, "DeepMIMO: A generic deep learning dataset for millimeter wave and massive MIMO applications," in *ITA*, San Diego, CA, 2019.

[22] Remcom, "Wireless InSite," http://www.remcom.com/wireless-insite.

[23] A. Vaswani, N. Shazeer, N. Parmar *et al.*, "Attention is all you need," in *NeurIPS*, 2017.

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[25] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.

A Dataset

We use the following scenarios for pretraining: ASU Campus, Boston, Amsterdam, Athens, Bangkok, Barcelona, Beijing, Brussels, Cairo, Cape Town, Charlotte, Chicago, Denver, Dubai, Edinburgh, Florence, Fort Worth, Fujiyoshida, Granada, Hatsukaichi, Helsinki, Hong Kong, Indianapolis, Istanbul, Jerusalem, Kyoto, Havana, Lisbon, Los Angeles, Madrid, Marrakesh, Mumbai, New Delhi, New York, North Jakarta, Oklahoma City, Philadelphia, Phoenix, Reykjavik, Rio de Janeiro, Rome, San Francisco, San Nicolas, San Francisco, Saint Petersburg, Santa Clara, Santiago, Seattle, Seoul, Singapore, Stockholm, Sumida City, Sydney, Taipei, Taito City, Toronto, and Warsaw.

For evaluation we use: Austin, Centro, Columbus, Dallas, Gurbchen, Houston, Miami, Montreal, Prague, and San Diego.