

Localização 3D de Equipamentos de Usuário via Fusão Multimodal Rádio-Visual em Redes mmWave

Wesley L. Passos^{1,2}, Jonathan N. Gois^{1,2} e José F. de Rezende²

Resumo—Este trabalho propõe uma arquitetura de fusão de dados multimodais para estimação de posição 3D de equipamentos de usuário em ambientes *indoor*, avaliada no conjunto de dados do CONVERGE Challenge Task 2. A arquitetura combina *features* visuais de um *backbone* ResNet-18, medidas de rádio da interface E2 em uma arquitetura O-RAN e um prior geométrico de detecção via YOLOv8. Três mecanismos de fusão de dados multimodais são avaliados. Um protocolo robusto de avaliação é utilizado para avaliar a capacidade de generalização dos modelos entre diferentes trajetórias do usuário. O melhor modelo atingiu uma redução de 14% sobre o estado da arte.

Palavras-Chave—Localização de UE, dados multimodais, O-RAN, aprendizado de máquina, visão computacional.

Abstract—This paper proposes a multimodal data fusion architecture for estimating the 3D position of user devices in indoor environments, evaluated on the CONVERGE Challenge Task 2 dataset. The architecture combines visual features from a ResNet-18 backbone, radio measurements from the O-RAN protocol's E2 interface, and a geometric detection prior via YOLOv8. Three multimodal data fusion mechanisms are evaluated. A robust evaluation protocol is used to assess the models' generalization ability across different user trajectories. The best model achieved a 14% improvement over the state of the art.

Keywords—UE localization, multimodal data, O-RAN, machine learning, computer vision.

I. INTRODUÇÃO

A localização precisa de equipamentos de usuário (UE) em ambientes *indoor* é um requisito importante para serviços de comunicação, onde a ausência de sinal de sistema global de navegação por satélite (GNSS) torna necessário o desenvolvimento de métodos alternativos [1]. Sistemas de rádio mmWave com arquitetura O-RAN, como os empregados na quinta geração de redes sem fio (5G NR), disponibilizam métricas detalhadas de canal pela interface E2, entre eles *reference signal received power* (RSRP), *reference signal received quality* (RSRQ), *signal-to-interference-plus-Noise ratio* (SINR), *received signal strength indicator* (RSSI) e *channel quality indicator* (CQI), que carregam informação implícita sobre a geometria da propagação e, portanto, sobre a posição do UE. Paralelamente, câmeras RGB-D integradas às estações base fornecem contexto visual rico do ambiente, complementando a resolução espacial limitada das medidas de rádio. A fusão dessas duas fontes de informação constitui uma direção promissora para localização de alta precisão, alinhada à visão

de sensoriamento integrado à comunicação (ISAC) em redes de próxima geração [2].

A localização *indoor* utilizando sinais de rádio é um tema bastante estudado na literatura. Revisões recentes [1], [3] documentam a evolução das abordagens baseadas em RSSI, de métodos clássicos como *K-nearest neighbors* (KNN) e *support vector machines* (SVM) para redes neurais profundas, que superam a variabilidade do sinal por meio de representações latentes. Deng et al. [4] combinam RSSI, diferença de intensidade de sinal e medidas estatísticas como curtose para treinar uma ResNet [5], reduzindo o erro médio de localização em 56% frente ao estado da arte em ambientes Wi-Fi. Zhang et al. [6] integram funções de ativação treináveis a uma arquitetura neural para extrair representações mais discriminativas de sequências de RSSI. Liu et al. [7] abordam o cenário de conjuntos de dados reduzidos com um filtro de partículas bidirecional que funde dados inerciais e de alcance sem mapa de rádio pré-construído.

No contexto de redes 5G mmWave, Kazemi et al. [8] propõem um *framework* de *fingerprinting* por CSI no lado do UE, com erro médio de 60 cm, utilizando redes multicamada que combinam *features* de CSI robustas a erros de sincronização. Kia et al. [9] empregam medidas CSI beamformadas de uma única estação base para posicionamento urbano via CNN, atingindo erro mínimo de 0,98 m em cenário *outdoor*. Esses trabalhos evidenciam o potencial do sinal de rádio 5G para localização, porém operam em modalidade única, sem fusão com informação visual.

A integração de câmeras RGB-D com sinais de rádio para localização de UE é explorada no CONVERGE Challenge Task 2 (CCT2) [2], cujo conjunto de dados multimodal fornece a base experimental deste trabalho. Saba et al. [10] propõem uma abordagem multimodal que concatena *embeddings* visuais de um *backbone* ResNet-50 congelado com cinco *features* E2 de rádio (1029-D) e os regide por uma MLP para posição 3D, obtendo MAE_{3D} de 51,9 mm. Contudo, dois aspectos limitam esse trabalho: a concatenação simples não modela a interação entre modalidades, impedindo que as medidas de rádio condicionem a interpretação das *features* visuais; e o protocolo de partição aleatória dos dados entre treinamento e teste cria sobreposição espacial entre treino e teste, inflando as métricas de desempenho.

No presente trabalho é proposta uma arquitetura de fusão rádio-visual treinável de ponta a ponta (*end-to-end*) para estimação de posição 3D de UE no conjunto de dados disponibilizado no CCT2 [2] com as seguintes contribuições:

- avaliação de três estratégias de fusão multimodal, incluindo as quais as *features* de rádio modulam adaptativamente as representações visuais;

¹CEFET-RJ, Rio de Janeiro-RJ. ²PESC/COPPE, UFRJ, Rio de Janeiro-RJ. e-mails: {wesley.passos, jonathan.gois}@cefet-rj.br, rezende@land.ufrj.br. Este trabalho foi financiado pelo Projeto Brasil 6G (MCTI/RNP 01245.020548/2021-07); pelo Projeto Smart 5G Core And MUltiRAN Integration (SAMURAI) (MCTIC/CGI.br/FAPESP 2020/05127-2); e pelo CNPq (403327/2025-3), através da Chamada Universal.



Fig. 1. Amostras de frames RGB da trajetória exp6, capturadas pela câmera Nerian Ruby RGB-D acoplada à estação base. O UE (Turtlebot 4) aparece em diferentes posições ao longo da trajetória.

- incorporação de mapa de disparidade (RGB-D) e prior geométrico de detecção via detector de objeto, ampliando a representação visual com informação estrutural e semântica; e
- protocolo de avaliação robusto, que mede a generalização a trajetórias não conhecidas, revelando limitações ocultas pelo protocolo de partição aleatória.

Para melhor apresentar essas contribuições o restante do presente trabalho está organizado da seguinte forma: a Seção II detalha a metodologia, incluindo o conjunto de dados, a arquitetura proposta e os protocolos de avaliação; a Seção III apresenta os resultados; e a Seção IV conclui o trabalho com um resumo dos achados e perspectivas para trabalhos futuros.

II. METODOLOGIA

A. Conjunto de dados

O conjunto de dados do CCT2 [2] é coletado em um ambiente *indoor* com uma estação base 5G mmWave (LiteOn FR2, pilha OAI) montada sobre um braço robótico UR-10e e um UE Quectel FR2 embarcado num robô Turtlebot 4. A Fig. 1 ilustra amostras do ambiente experimental capturadas pela câmera da estação base. O conjunto de dados contém três trajetórias (exp6, exp7, exp8) gravadas de forma sincronizada via PTP com: (i) câmera *stereo* Nerian Ruby RGB-D acoplada à estação base, fornecendo quadros RGB e mapas de disparidade a 7 fps (142 ms por quadro); (ii) métricas de interface E2 do protocolo O-RAN (como RSRP, RSRQ, SINR, RSSI e CQI) amostradas a cada 50 ms; e (iii) posição 3D de referência do UE obtida pelo sistema de captura de movimento Qualisys com precisão sub-milimétrica (*ground truth* em mm). Nas trajetórias exp6 e exp7 o UE permanece dentro do campo de visão (FoV, do inglês *field of view*) da câmera; e na trajetória exp8 o robô inicia e termina fora do FoV.

B. Modelo proposto

A Fig. 2 ilustra o pipeline do modelo proposto, que é composto por três módulos: extração de *features*; fusão de *features*; e regressão.

1) *Extração de features*: A extração de *features* é dividida em 3 ramos: (i) visuais, (ii) rádio (E2) e (iii) geométricas. Para extração de *features* visuais é utilizado um *backbone* ResNet-18 [5], com pesos pré-treinados na base ImageNet. Esse *backbone* processa o quadro de entrada (RGB, RGB-D ou disparidade), com a primeira camada convolucional adaptada ao número de canais da entrada. No caso do canal de disparidade, a camada é inicializada como a média dos pesos RGB para aproveitar as representações pré-treinadas.

A camada de classificação é substituída por uma identidade, produzindo um vetor de *features* visuais $\mathbf{v} \in \mathbb{R}^{512}$.

As *features* de rádio são extraídas das seguintes medidas E2: RSRP, RSRQ, SINR, RSSI, e CQI, as quais são normalizadas via *z-score* e passam por uma MLP de dois estágios ($5 \rightarrow 64 \rightarrow 64$, ReLU, e *dropout* com probabilidade $p_d=0,1$), produzindo uma representação densa $\mathbf{r} \in \mathbb{R}^{64}$.

Para injetar um prior geométrico explícito, é anexado um descritor de detecção compacto extraído de um modelo YOLOv8 [11] pré-treinado que detecta a posição do robô na imagem. O resultado de cada quadro é compactado em um vetor $\mathbf{b} = [x_c; y_c; w; h; s; d] \in \mathbb{R}^6$, onde x_c e y_c são as coordenadas do centro da *bounding box* na imagem, w e h a largura e altura da *bounding box*, $s \in [0, 1]$ é o nível de confiança da detecção pelo modelo YOLO e d um indicador binário de detecção válida. Quando não há detecção, $(x_c; y_c)$ assume o valor padrão (0,5; 0,5) e o indicador $d = 0$.

2) *Mecanismo de fusão*: Dadas as representações $\mathbf{v} \in \mathbb{R}^{512}$, $\mathbf{r} \in \mathbb{R}^{64}$ e $\mathbf{b} \in \mathbb{R}^6$, três estratégias de fusão são avaliadas: concatenação simples (concat.), portão por modalidade (GMU, do inglês *gated multimodal unit*) [12] e modulação linear por feature (FiLM, do inglês *feature-wise linear modulation*) [13].

Na fusão por concatenação simples, as representações são simplesmente concatenadas em um vetor $\mathbf{f} = [\mathbf{v}; \mathbf{r}; \mathbf{b}]$.

O mecanismo GMU [12] aprende a ponderar a contribuição de cada modalidade por meio de portões multiplicativos diferenciáveis. Em nossa adaptação, o vetor de rádio $\mathbf{r}$ condiciona um portão vetorial $\mathbf{g}$ que modula por canal a representação visual $\mathbf{v}$:

$$\mathbf{g} = \sigma(W_g \mathbf{r} + \beta_g) \in [0, 1]^{512}, \quad (1)$$

$$\tilde{\mathbf{v}}_g = \mathbf{g} \odot \mathbf{v}, \quad (2)$$

onde $W_g \in \mathbb{R}^{512 \times 64}$ é a matriz de projeção aprendida, $\beta_g \in \mathbb{R}^{512}$ é o vetor de vies, $\sigma(\cdot)$ é a função sigmoide aplicada elemento a elemento e $\odot$ denota o produto de Hadamard. Cada componente $g_i \in [0, 1]$ regula individualmente a transmissão do feature visual v_i : valores próximos de zero suprimem a feature, enquanto valores próximos de um a transmitem integralmente. O vetor modulado $\tilde{\mathbf{v}}_g \in \mathbb{R}^{512}$ substitui $\mathbf{v}$ na concatenação final $\mathbf{f} = [\tilde{\mathbf{v}}_g; \mathbf{r}; \mathbf{b}]$.

O mecanismo FiLM [13] modula as *features* visuais por meio de uma transformação afim condicionada pela representação de rádio $\mathbf{r}$. Uma única projeção linear gera simultaneamente os parâmetros de escala $\gamma \in \mathbb{R}^{512}$ e translação $\sigma \in \mathbb{R}^{512}$:

$$[\gamma; \sigma] = W_f \mathbf{r} + \beta_f, \quad (3)$$

$$\tilde{\mathbf{v}}_f = \mathbf{v} \odot (1 + \gamma) + \sigma, \quad (4)$$

onde $W_f \in \mathbb{R}^{1024 \times 64}$ e $\beta_f \in \mathbb{R}^{1024}$ são parâmetros aprendidos e $\odot$ denota o produto de Hadamard. A formulação residual $(1 + \gamma)$ inicializa o comportamento próximo à identidade ($\gamma \approx 0$ no início do treinamento), favorecendo a estabilidade com conjuntos de dados pequenos. Por não restringir γ ao intervalo $[0, 1]$, ao contrário do portão sigmoide do GMU, o FiLM permite supressão ($\gamma_i \approx -1$), amplificação ($|\gamma_i| > 1$) e inversão de sinal ($\gamma_i < -1$) de *features* individuais, conferindo maior expressividade ao condicionamento rádio-visual.

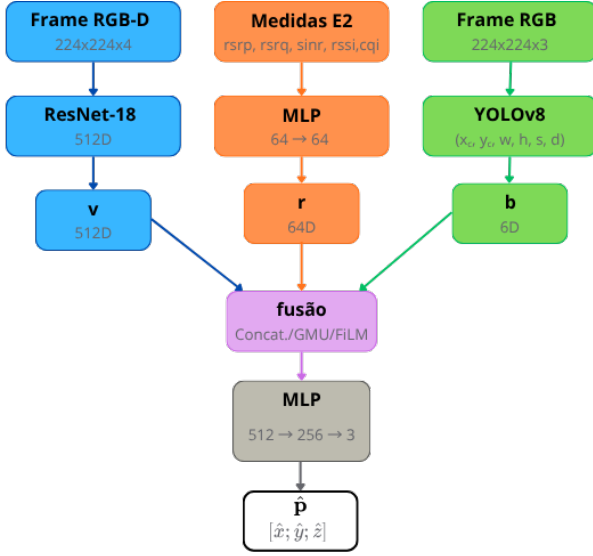


Fig. 2. Arquitetura proposta. Três ramos extraem representações complementares: (i) visual: backbone ResNet-18 adaptado para entrada RGB-D produz $\mathbf{v} \in \mathbb{R}^{512}$; (ii) rádio: MLP aplicada às cinco medidas E2 produz $\mathbf{r} \in \mathbb{R}^{64}$; e (iii) geométrico: detecção do modelo YOLOv8 compactado em $\mathbf{b} \in \mathbb{R}^6$ (centro, dimensões, confiança e validade da *bounding box*). As representações são fundidas por um dos mecanismos avaliados (Concat., GMU ou FiLM) e regredidas para a posição 3D $\hat{\mathbf{p}} = [\hat{x}; \hat{y}; \hat{z}]$ em mm.

O vetor modulado $\tilde{\mathbf{v}}_f \in \mathbb{R}^{512}$ substitui $\mathbf{v}$ na concatenação final $\mathbf{f} = [\tilde{\mathbf{v}}_f; \mathbf{r}; \mathbf{b}]$.

3) *Regressão*: Após a fusão das representações por algum dos mecanismos descritos, uma rede MLP com duas camadas ocultas (512 e 256 unidades, ReLU, *dropout* $p=0,1$) realiza a regressão para estimar as coordenadas 3D (x, y, z) .

C. Métricas e protocolos de avaliação

No presente trabalho, o erro de localização 3D é definido como a distância euclidiana 3D entre as posições previstas e as medidas. As métricas de avaliação são o erro absoluto médio (MAE_{3D}) e a raiz do erro quadrático médio (RMSE_{3D}), calculadas respectivamente como

$$\text{MAE}_{3D} = \frac{1}{N} \sum_{i=1}^N \|\hat{\mathbf{p}}_i - \mathbf{p}_i\| \quad \text{e} \quad (5)$$

$$\text{RMSE}_{3D} = \sqrt{\frac{1}{N} \sum_{i=1}^N \|\hat{\mathbf{p}}_i - \mathbf{p}_i\|^2}, \quad (6)$$

onde $\hat{\mathbf{p}} = [\hat{x}; \hat{y}; \hat{z}]$ e $\mathbf{p} = [x; y; z]$ são, respectivamente, a predição e a anotação da posição 3D da UE.

Dois protocolos de avaliação são empregados: (i) *partição aleatória* (80/20): esse protocolo é o mesmo aplicado em Saba et al. [10]. Neste protocolo, as amostras são divididas aleatoriamente em 80% para treinamento e 20% para teste; e (ii) *avaliação inter-trajetória* (LOTO, do inglês *leave-one-trajectory-out*): treinamento em duas trajetórias e avaliação na trajetória restante (*holdout 3-fold*), utilizada para medir a generalização a trajetórias não conhecidas durante o treinamento.

D. Configurações experimentais

Todos os modelos com *backbone* treinável (ResNet-18) são otimizados utilizando Adam, com taxa de aprendizado

$\alpha = 10^{-4}$ e *weight decay* $\lambda_{\text{wd}} = 10^{-4}$, por 100 épocas com *batch size* 16. A função de perda empregada é a função de Huber [14], dada por

$$L_{\delta}(t, \hat{t}) = \begin{cases} \frac{1}{2}(t - \hat{t})^2, & \text{para } |t - \hat{t}| \leq \delta \\ \delta \left(|t - \hat{t}| - \frac{1}{2}\delta \right), & \text{caso contrário,} \end{cases} \quad (7)$$

onde t e $\hat{t}$ são, respectivamente, os valores observados e estimados, e $\delta=1,0$ especifica o limite a partir do qual se alterna entre as perdas ℓ_1 e ℓ_2 . O alvo é normalizado com *z-score* calculado sobre o conjunto de treino de cada *fold*. O modelo com o melhor MAE_{3D} de validação é selecionado para teste.

1) *Congelamento dos backbones*: Para conjuntos de dados de localização pequenos, o ajuste fino completo do backbone pode causar sobreajuste. Portanto, nos experimentos, três regimes são avaliados: (i) sem congelamento: todos os parâmetros são treináveis; (ii) congelamento parcial: as camadas 1 e 2 são congeladas (mantidas em modo *eval* para preservar as estatísticas do *BatchNorm* pré-treinado), enquanto as camadas 3 e 4, juntamente com o regressor, são ajustados; e (iii) congelamento total: todo o *backbone* é congelado e somente o regressor é treinado.

2) *Baseline*: Dois métodos são utilizados como *baseline* para fins de comparação com o método proposto. O primeiro, reportado em Saba et al. [10], utiliza *features* extraídas de um *backbone* ResNet-50-L3 (saída da camada 3 – L3) congelado, de dimensão 1024, concatenadas às *features* de rádio *z-scored*, de dimensão 5, totalizando um vetor de dimensão 1029. O vetor de features resultante (1029D) é aplicado em uma MLP de quatro camadas (512, 256, 128, 64) que realiza regressão para predição da posição 3D da UE.

Para verificar a conjectura de que no protocolo de partição aleatória, a MLP de Saba et al. [10] aprende implicitamente a interpolar a média dos vizinhos mais próximos no espaço de *embedding* em vez de aprender uma correspondência posição–aparência generalizável, propõe-se substituir a MLP por um regressor KNN, mantendo os mesmos *embeddings*. Três variantes desse experimento são avaliadas: (i) rádio (E2): apenas as cinco *features* de rádio normalizadas; (ii) visual: apenas o *embedding* ResNet-50-L3 (1024-D); e (iii) visual+E2: concatenação dos dois vetores (1029-D).

III. RESULTADOS E DISCUSSÃO

A. Partição aleatória (80/20) e ablação de componentes

Na Tabela I, são apresentados os resultados dos modelos no protocolo de partição aleatória (80/20), em três grupos de ablação. Neste protocolo, quadros da mesma trajetória distribuem-se entre treino e teste, criando sobreposição espacial implícita que favorece métodos baseados em similaridade de representações. Assim, os resultados devem ser interpretados como limite superior de desempenho.

Com *features* extraídas da ResNet-50-L3 sem ajuste fino, o regressor KNN com $k=3$ atingiu $\text{MAE}_{3D}=35,7$ mm, superando o resultado reportado por Saba et al. [10] sem qualquer treinamento de parâmetros. A adição de *features* de rádio E2 (combinação visual+E2), resulta em $\text{MAE}=34,0$ mm

TABELA I

RESULTADOS DE MAE_{3D} E RMSE_{3D}, PARTIÇÃO ALEATÓRIA (80/20).
BACKBONE: RESNET-50-L3 PARA KNN; RESNET-18 TREINÁVEL PARA AS
DEMAIS. VALORES EM MM. *VALOR REPORTADO ORIGINALMENTE PELOS
AUTORES, CALCULADO COMO MÉDIA DOS ERROS POR COORDENADA.

Método	MAE _{3D}	RMSE _{3D}
Saba et al. [10]*	51,9	56,3
<i>Backbone congelado (ResNet-50-L3)</i>		
KNN ($k=1$), E2	621,3	985,3
KNN ($k=3$), visual	35,7	69,0
KNN ($k=3$), visual + E2	34,0	67,5
<i>Ablação canal visual e YOLO — Concat., sem congelamento</i>		
Concat. RGBD+E2 (sem YOLO)	123,2	197,1
Concat. RGBD+E2+YOLO	111,4	182,3
Concat. RGB+E2+YOLO	111,0	160,2
<i>Ablação fusão e congelado — RGBD+E2+YOLO</i>		
Concat. (congelamento parcial)	111,4	186,1
Concat. (congelamento total)	255,7	344,4
GMU (sem congelamento)	170,5	217,7
GMU (congelamento parcial)	94,7	174,2
GMU (congelamento total)	269,0	348,2
FiLM (sem congelamento)	91,5	165,4
Apenas E2 (rádio)	952,5	1093,8

($\Delta=1,7$ mm, $\approx 5\%$), superando o KNN puramente visual, sendo a melhor configuração deste grupo. Esses resultados evidenciam que, na partição aleatória, a informação visual captura a correlação posição-aparência de forma mais direta do que um MLP de regressão, aproveitando a sobreposição espacial treino/teste. Como as amostras pertencem a trajetórias contínuas, a partição aleatória pode colocar no treino quadros temporal e espacialmente muito próximos dos quadros de teste. Assim, o bom desempenho do KNN sugere que o método explora a proximidade entre *embeddings* de cenas quase equivalentes, funcionando mais como um diagnóstico de sobreposição espacial do que como evidência de generalização.

Vale ressaltar que, no presente trabalho, o erro de localização 3D é reportado como a distância euclidiana entre as posições previstas e as posições reais, enquanto Saba et al. [10] reportam os erros calculados como a média das componentes das coordenadas. Portanto, deve-se observar que os erros euclidianos 3D são naturalmente maiores do que os erros médios por coordenada e, assim, as duas formulações não são diretamente comparáveis.

Tomando como base a variante ‘concat. RGBD+E2’ sem YOLO (123,2 mm), a inclusão do *bounding box* do robô via YOLO reduziu o MAE em $\approx 10\%$ (111,4 mm), confirmando que a localização semântica do objeto na imagem é informação complementar relevante. A substituição do canal de profundidade (RGBD $\rightarrow$ RGB) não alterou o MAE_{3D} (111,0 vs. 111,4 mm), mas reduziu o RMSE_{3D} de 182,3 para 160,2 mm, indicando que a profundidade introduziu ruído.

Entre os mecanismos de fusão sem congelamento, o FiLM obteve o menor MAE_{3D} (91,5 mm), seguido de concat. (111,4 mm) e GMU (170,5 mm). O congelamento das camadas 1–2 beneficia especialmente o mecanismo GMU, reduzindo seu MAE em 44% (170,5 $\rightarrow$ 94,7 mm) e aproximando-o do

FiLM sem congelamento, enquanto o mesmo regime tem efeito neutro sobre o concat. (111,4 mm em ambos os casos). O congelamento total degrada todas as variantes (concat.: 255,7 mm; GMU: 269,0 mm), confirmando que a adaptação do *backbone* à tarefa é essencial. O comportamento do GMU sugere que seu mecanismo de portão é mais sensível à qualidade das representações visuais intermediárias, que são estabilizadas pelo congelamento parcial das camadas de baixo nível. Em conjuntos de dados pequenos, o congelamento total impede a adaptação das representações à tarefa, enquanto o ajuste fino completo superajusta. O congelamento parcial equilibra generalização e especialização, sendo a estratégia recomendada quando o conjunto de dados de localização é limitado.

B. Avaliação inter-trajetória (LOTO)

Na Tabela II, são apresentados os resultados do protocolo de avaliação inter-trajetória (LOTO). O KNN visual, que atingiu 35,7 mm na partição aleatória (80/20), passa a 616 mm no LOTO, um aumento de $17\times$. Em trajetórias diferentes, o vizinho mais próximo no espaço de *embeddings* raramente corresponde a uma posição próxima, pois as representações de um *backbone* congelado são sensíveis a condições de iluminação, disposição de objetos e perspectiva específicas. O método de Saba et al. [10] (replicado), embora também utilize *backbone* congelado, beneficia-se do ajuste fino da MLP de regressão, alcançando MAE médio de 537 mm.

Todos os modelos com *backbone* ResNet-18 treinável reduzem consistentemente o MAE_{3D} em relação às referências com *backbone* congelado: a variante mais simples, ‘concat. RGBD+E2 (sem YOLO)’, atingiu 467 mm, uma redução de 13% sobre o MLP Saba (537 mm). O ajuste fino *end-to-end* permite ao *backbone* aprender variações específicas à tarefa, melhorando a transferência entre trajetórias.

O modelo ‘FiLM RGBD+E2+YOLO’ obteve o menor MAE_{3D} médio: 461 mm (RMSE = 521 mm), com desempenho consistente em exp6 (172,7 mm) e exp7 (169,4 mm). A modulação FiLM [13], que condiciona as ativações visuais ao vetor de rádio por escalonamento e translação afim, favorece a generalização ao adaptar a representação visual à condição de canal observada.

O ‘GMU RGBD+E2+YOLO’ apresentou o menor MAE em exp6 (166,5 mm), porém alta variância entre *folds* (372,2 mm em exp7), com MAE médio de 515 mm. Apesar de ser a melhor configuração na partição aleatória com congelamento parcial (Tabela I), esse mecanismo não generaliza para trajetórias desconhecidas com a mesma eficácia.

No LOTO, o impacto da posição do *bounding box* é marginal. A configuração ‘concat. RGBD+E2’ (467 mm) e ‘concat. RGBD+E2+YOLO’ (470 mm) são equivalentes, ao contrário dos 10% de ganho observados no protocolo que utiliza partição aleatória (80/20) dos dados. A razão RMSE/MAE situa-se entre 1,10 e 1,16 nos *folds* exp6 e exp7, indicando distribuições de erro aproximadamente simétricas. No *fold* correspondente à trajetória exp8, erros extremos associados ao robô fora do campo visual inflam o RMSE desproporcionalmente.

TABELA II

RESULTADOS DE MAE_{3D} E RMSE_{3D} POR *fold holdout* E MÉDIA $\pm$ DESVIO PADRÃO, INTER-TRAJETÓRIA (LOTO). VALORES EM MM. LINHAS EM ITÁLICO: REFERÊNCIAS SEM AJUSTE FINO END-TO-END. DEMAIS: RESNET-18 TREINÁVEL. [†]EM *exp8* O ROBÔ SAI DO FoV EM 14,7% DOS FRAMES.

Método	<i>exp6</i>		<i>exp7</i>		<i>exp8</i> [†]		Média $\pm$ std	
	MAE _{3D}	RMSE _{3D}	MAE _{3D}	RMSE _{3D}	MAE _{3D}	RMSE _{3D}	MAE _{3D}	RMSE _{3D}
<i>Saba et al. [10] (replicação)</i>	209,6	241,6	190,0	216,4	1211,7	1365,8	537 $\pm$ 477	608 $\pm$ 536
<i>KNN, visual</i>	251,7	290,1	373,0	429,3	1224,5	1447,7	616 $\pm$ 433	722 $\pm$ 516
Concat. RGBD+E2	178,9	216,4	179,4	195,2	1042,8	1163,4	467 $\pm$ 407	525 $\pm$ 452
Concat. RGBD+E2+YOLO	221,5	240,4	154,0	171,7	1035,1	1153,6	470 $\pm$ 400	522 $\pm$ 448
Concat. RGB+E2+YOLO	204,3	233,9	194,5	224,0	1053,2	1223,2	484 $\pm$ 403	560 $\pm$ 469
GMU RGBD+E2+YOLO	166,5	184,9	372,2	380,7	1006,0	1141,6	515 $\pm$ 357	569 $\pm$ 413
FiLM RGBD+E2+YOLO	172,7	194,1	169,4	185,4	1040,7	1183,8	461$\pm$410	521$\pm$469
Concat. Disparidade+E2+YOLO	237,0	266,9	230,2	279,4	922,6	1002,3	463 $\pm$ 325	516 $\pm$ 344
Apenas E2 (rádio)	759,5	867,1	956,7	1019,1	3581,8	3861,1	1766 $\pm$ 1287	1916 $\pm$ 1377

A variante ‘concat. Disparidade+E2+YOLO’ alcançou 922,6 mm de MAE_{3D} em *exp8* (RMSE=1002,3 mm), reduzindo o erro em $\approx$ 118 mm frente ao FiLM e apresentando o menor desvio padrão médio (325 mm vs. $\approx$ 400 mm dos demais). A disparidade codifica a estrutura geométrica do ambiente ao redor da última posição detectada do robô, oferecendo pistas de localização residuais mesmo quando a aparência visual do objeto está ausente.

A transição do protocolo de partição aleatória (80/20) para o inter-trajetória (LOTO) revela o efeito real do *overfitting* à distribuição das trajetórias de treinamento. Métodos que dependem de sobreposição espacial colapsam, enquanto modelos com representações visuais aprendidas *end-to-end* apresentam maior resiliência.

Especificamente na trajetória *exp8*, o robô sai do campo visual da câmera em 14,7% dos frames. Todos os métodos predominantemente visuais colapsam, apresentando erros que superam em mais de uma ordem de grandeza os obtidos em *exp6* e *exp7*. A degradação na trajetória *exp8* reflete uma limitação estrutural inerente ao conjunto de dados. Posições fora do FoV correspondem a pontos de operação fora da distribuição de treino, independentemente da trajetória de *holdout*. Uma direção promissora é a fusão adaptativa de *features*, de modo que, quando a confiança de detecção do modelo YOLO estiver próxima de zero, o sistema poderia atribuir maior peso à modalidade de rádio, que permanece sempre disponível.

IV. CONCLUSÃO

Este trabalho propôs uma arquitetura de fusão rádio-visual treinável de ponta a ponta para estimação de posição 3D de UE, avaliada no conjunto de dados do CONVERGE *Challenge Task 2*. A fusão FiLM com backbone ResNet-18 e prior geométrica de detecção via YOLOv8 obteve MAE_{3D} médio de 461 mm no protocolo LOTO, redução de 14% sobre de Saba et al. (537 mm). O protocolo LOTO revelou que a partição aleatória (80/20) infla artificialmente o desempenho por sobreposição espacial entre treinamento e teste. O mecanismo de fusão FiLM mostrou maior consistência inter-trajetória que o GMU, e o congelamento parcial do *backbone* mostrou-se a melhor estratégia para o conjunto de dados reduzido. A

principal limitação é a degradação quando o UE sai do campo de visão, cenário em que a variante com disparidade mostrou-se a mais robusta. Como trabalhos futuros, destacam-se a fusão adaptativa de modalidades, a agregação de informação temporal entre amostras consecutivas e a avaliação com sinais de referência de sondagem (SRS).

REFERÊNCIAS

- [1] B. Zholamanov, A. Saymbetov, M. Nurgaliyev, A. Bolatbek, G. Dosymbetova, N. Kuttybay, S. Orynassar, A. Kapparova, N. Koshkarbay, and Ö. F. Beyca, “RSSI fingerprint-based indoor localization solutions using machine learning algorithms: A comprehensive review,” *Smart Cities*, vol. 8, no. 5, p. 153, 2025.
- [2] J. Chen, F. B. Teixeira, F. M. Ribeiro, A. Alkhateeb, M. Ricardo, L. M. Pessoa, and D. Stock, “Converge Challenge: Multimodal Learning for 6G Wireless Communications,” in *IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 21799–21801, 2026.
- [3] B. Chen, J. Ma, L. Zhang, J. Zhou, J. Fan, and H. Lan, “Research progress of wireless positioning methods based on RSSI,” *Electronics*, vol. 13, no. 2, p. 360, 2024.
- [4] S. Deng, W. Zhang, L. Xu, and J. Yang, “RRIFLoc: Radio robust image fingerprint indoor localization algorithm based on deep residual networks,” *IEEE Sens. J.*, vol. 23, no. 3, pp. 3233–3242, 2023.
- [5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 770–778, 2016.
- [6] S. Zhang, W. Zhang, R. Chen, G. Pan, C. Hu, and H. Li, “AFLoc: A deep learning indoor localization method with trainable activation functions,” *IEEE Internet Things J.*, vol. 13, no. 4, pp. 6343–6354, 2026.
- [7] X. Liu, Y. Zhao, X. Li, and C.-Z. Xu, “Bi-directional particle filter fusion for fingerprinting indoor localization with few short tracks,” *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–15, 2025.
- [8] P. Kazemi, H. Al-Tous, C. Studer, and O. Tirkkonen, “User-side indoor localization using CSI fingerprinting,” in *IEEE Int. Workshop Signal Process. Adv. Wirel. Commun.*, pp. 1–5, 2022.
- [9] G. Kia, L. Ruotsalainen, and J. Talvitie, “A CNN approach for 5G Mm wave positioning using beamformed csi measurements,” in *Int. Conf. Loc. GNSS*, pp. 01–07, 2022.
- [10] N. Saba and H. Al-Tous, “ICASSP 2026 CONVERGE Challenge: Multimodal Fingerprinting for UE Localization,” in *IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 21802–21803, 2026.
- [11] J. Redmon, “You only look once: Unified, real-time object detection,” in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 779–788, 2016.
- [12] J. Arevalo, T. Solorio, M. M. y Gómez, and F. A. González, “Gated multimodal networks,” *Neural Comput. Appl.*, 2020.
- [13] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, “FiLM: Visual reasoning with a general conditioning layer,” in *Proc. AAAI Conf. Artif. Intell.*, 2018.
- [14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 ed., 2009.