ST-MTLNet: Representações Espaço-Temporais de Pontos de Interesse para Aprendizado Multitarefa

Tarik S. Paiva[1], **Vitor H. O. Silva**[1], **Germano B. dos Santos**[1], **Fabrício A. Silva**[1]

[1]Laboratório de Inteligência em Sistemas Pervasivos e Distribuídos (NESPeD-LAB)
Universidade Federal de Viçosa, Florestal, MG, Brasil

{tarik.paiva, vitor.oliveira, germano.santos, fabricio.asilva}@ufv.br

Abstract. *This work proposes ST-MTLNet, a multi-task architecture for POI Category Classification and Next-POI Prediction based on decoupled representations. The model combines a continuous spatial encoder for geographic coordinates, a temporal encoder (Time2Vec) for visitation patterns, and a hierarchical categorical encoder (HGI) for structural and regional POI context. Two spatial encoders, SIREN and Sphere2Vec-M, originally proposed for remote sensing and ecology tasks, are evaluated in the context of multi-task POI modeling on LBSNs. Experiments on the Gowalla dataset across Florida, California, and Texas show that the proposed approach outperforms the baseline in all 21 category-state combinations for classification, with average gains of 20–24 percentage points, and in 76% of combinations for next-POI prediction. The comparison between spatial encoders also reveals complementary performance profiles associated with the geographic distribution of POIs in each territory.*

Resumo. *Este trabalho propõe o ST-MTLNet, uma arquitetura multitarefa para classificação de categoria de POI e predição do próximo POI baseada em representações desacopladas. O modelo combina uma representação espacial contínua para coordenadas geográficas, uma representação temporal (Time2Vec) para padrões de visitação e uma representação categórica hierárquica (HGI) para contexto estrutural e regional dos POIs. Duas arquiteturas de codificação espacial, SIREN e Sphere2Vec-M, originalmente propostas para sensoriamento remoto e ecologia, são avaliadas no contexto de tarefas multitarefa de POIs em LBSNs. Experimentos com o dataset Gowalla nos estados da Flórida, Califórnia e Texas demonstram que a abordagem proposta supera o baseline em todas as 21 combinações de categoria e estado para classificação, com ganhos médios de 20 a 24 pontos percentuais, e em 76% das combinações para predição do próximo POI. A comparação entre as arquiteturas espaciais revela ainda perfis complementares de desempenho associados à distribuição geográfica dos POIs em cada território.*

1. Introdução

Redes Sociais Baseadas em Localização (*Location-Based Social Networks*, LBSNs), como o Gowalla [Cho et al. 2011, Jure 2014], registram *check-ins* de usuários com localização geográfica, horário de visita e categoria do local visitado. Nesse contexto, um Ponto de Interesse (POI) corresponde a um local físico semanticamente identificável, como restaurantes, lojas, aeroportos ou parques. A análise desses dados é relevante para aplicações como recomendação e mobilidade urbana. Nesse cenário, duas tarefas se destacam:

1. **Classificação de Categoria de POI**: classificar a categoria semântica (e.g., *Food*, *Shopping*, *Travel*) de um POI a partir de suas características.

2. **Predição do Próximo POI**: prever a categoria do próximo POI que um usuário visitará, dada a sequência histórica de seus *check-ins*.

Em princípio, essas tarefas podem ser tratadas como relacionadas, pois compartilham a mesma base de dados de mobilidade e o mesmo espaço semântico de categorias, o que torna o aprendizado multitarefa (*Multitask Learning*, MTL) uma estratégia promissora para explorar informação compartilhada entre elas [Caruana 1997]. No entanto, elas impõem demandas distintas. A classificação de categoria de POI é uma tarefa não sequencial, pois depende principalmente dos atributos do próprio POI e de seu contexto espacial. Já a predição do próximo POI é explicitamente sequencial, exigindo a modelagem de ordem temporal, recorrência e transições ao longo da trajetória do usuário. Assim, as informações úteis para uma tarefa podem não beneficiar a outra da mesma forma.

O MTLNet, proposto em [Silva et al. 2025], explora esse potencial ao modelar as duas tarefas em uma única arquitetura multitarefa com compartilhamento rígido de parâmetros. No entanto, o modelo utiliza exclusivamente representações vetoriais baseadas na estrutura do grafo, geradas pelo *Deep Graph Infomax* (DGI), que codifica informações espaciais e categóricas em um único vetor de 64 dimensões, sem incorporar explicitamente coordenadas geográficas contínuas nem padrões temporais de visitação.

Diante disso, surge o questionamento se a decomposição da entrada em *encoders* especializados para espaço, tempo e categoria produz uma base mais adequada para as duas tarefas do que o *embedding* monolítico do DGI. A hipótese é que representações desacopladas capturam dimensões complementares da mobilidade humana que não são modeladas explicitamente pela abordagem original.

Nesse sentido, este trabalho propõe o ST-MTLNet (*Spatial-Temporal MTLNet*), que integra três representações independentes: uma representação espacial contínua para coordenadas geográficas, uma representação temporal (Time2Vec) para padrões de visitação e uma representação categórica hierárquica para relações regionais e estruturais entre POIs. Na dimensão espacial, avaliamos duas arquiteturas com pressupostos distintos, SIREN e Sphere2Vec-M, originalmente validadas em tarefas geoespaciais de sensoriamento remoto e ecologia [Wu et al. 2024], mas ainda pouco exploradas no aprendizado multitarefa de POIs em LBSNs.

Os experimentos foram realizados com a base pública Gowalla [Jure 2014] nos estados da Flórida, Califórnia e Texas. Na classificação categórica, o ST-MTLNet supera o MTLNet em todas as combinações de categoria e estado avaliadas, com ganhos médios de 20 a 24 pontos percentuais. Na predição do próximo POI, o modelo proposto vence na maioria dos cenários, com destaque para categorias como *Food*. A comparação entre SIREN e Sphere2Vec-M mostra ainda que, embora a abordagem modular seja consistentemente superior ao *baseline*, o *encoder* espacial mais adequado depende da distribuição geográfica dos POIs em cada território.

Portanto, as principais contribuições deste trabalho são:

- A proposição do ST-MTLNet, que incorpora representações espaciais contínuas, temporais e categóricas hierárquicas ao MTLNet.
- A avaliação comparativa de SIREN e Sphere2Vec-M em três estados com estruturas territoriais distintas.
- A disponibilização pública do código-fonte e do *pipeline* experimental.[1]

[1]O código-fonte completo está disponível em `https://github.com/TarikSalles/Spatial_Embeddings`

A organização deste trabalho segue o seguinte formato: a Seção 2 apresenta os trabalhos relacionados. A Seção 3 descreve a metodologia proposta. A Seção 4 discute os resultados experimentais. Por fim, a Seção 5 apresenta as conclusões e trabalhos futuros.

2. Trabalhos Relacionados

A predição e classificação de POIs são desafios fundamentais em sistemas de recomendação baseados em localização e em análise de mobilidade urbana. Esta seção organiza os trabalhos relacionados em quatro eixos: representações baseadas em grafos, codificadores espaciais, representações temporais e aprendizado multitarefa aplicado a POIs.

2.1. Representações de POIs Baseadas em Grafos

A modelagem de POIs por meio de grafos permite capturar relações de vizinhança espacial e semântica entre locais. O modelo POI2Vec [Feng et al. 2017] adapta a arquitetura Word2Vec para gerar *embeddings* de POIs que incorporam a influência geográfica entre locais por meio de uma estrutura de árvore binária geográfica. O CATAPE (*Category-Aware Location Embedding*) [Rahmani et al. 2019] estende essa ideia incorporando informações categóricas e sequenciais, capturando a influência geográfica entre POIs com base na sequência temporal de visitas dos usuários.

Em uma abordagem auto-supervisionada, o *Deep Graph Infomax* (DGI) [Veličković et al. 2019] maximiza a informação mútua entre representações locais (nós) e globais (grafo) para gerar *embeddings* sem supervisão explícita. O HGI (*Hierarchical Graph Infomax*) [Huang et al. 2023] avança sobre essa linha ao operar em múltiplos níveis hierárquicos (POI, região e cidade), produzindo representações enriquecidas com informações regionais por meio de convoluções em grafos e mecanismos de multi-atenção.

No entanto, essas abordagens codificam informações espaciais de forma implícita por meio da topologia do grafo, sem utilizar diretamente as coordenadas geográficas como sinal de entrada para a representação.

2.2. Representações Espaciais

O uso direto de coordenadas geográficas para a geração de características latentes espaciais é uma abordagem distinta das representações baseadas em grafos. O *framework* TorchSpatial [Wu et al. 2024] estabelece um *benchmark* unificado para a avaliação de estratégias de representação espacial, permitindo a comparação sistemática entre diferentes arquiteturas.

A metodologia SIREN (*Sinusoidal Representation Networks*) [Sitzmann et al. 2020], aplicada ao contexto geográfico [Rußwurm et al. 2024], modela funções contínuas por meio de ativações senoidais, sendo capaz de representar sinais de alta frequência em dados espaciais. O Sphere2Vec [Mai et al. 2023] opera diretamente sobre coordenadas esféricas com múltiplas escalas de resolução, preservando propriedades de distância geodésica que são perdidas em projeções planares.

Embora essas estratégias tenham sido avaliadas em tarefas geoespaciais como predição de espécies e estimativa de população, sua aplicação pode ser estendida para modelos de predição e classificação de POIs como proposto neste trabalho.

2.3. Representações Temporais

A dimensão temporal é central para tarefas de predição de POIs, uma vez que os padrões de visitação dos usuários apresentam regularidades (e.g., horários de refeição, deslocamentos

semanais). O modelo LSTPM (*Long- and Short-Term Preference Modeling*) [Sun et al. 2020] captura preferências de curto e longo prazo utilizando uma Geo-Dilated RNN que relaciona POIs não consecutivos por meio de sequências temporais de visita e informações geográficas. O TransTARec [Sun 2024] propõe um modelo de tradução temporal adaptativa que une representações do POI, do *timestamp* e das preferências do usuário para predição do próximo POI.

Esses modelos incorporam a dimensão temporal principalmente no contexto da sequência de visitas do usuário. Nesse sentido, o Time2Vec [Kazemi et al. 2019] propõe uma representação temporal que combina termos lineares com funções senoidais de frequência e fase aprendíveis, capturando padrões periódicos e não periódicos a partir de valores temporais contínuos.

2.4. Aprendizado Multitarefa para Tarefas de POI

O aprendizado multitarefa (*Multitask Learning*, MTL) [Caruana 1997] tem sido explorado em tarefas de POIs principalmente para combinar a predição do próximo POI com sinais temporais, comportamentais ou regionais. O MCARNN [Liao et al. 2018] integra dependências sequenciais e regularidades temporais para predizer simultaneamente atividades e locais futuros. O MTPR [Xia et al. 2020] modela conjuntamente localização e contexto temporal por meio de LSTMs geográficas e aprendizado adversarial. Em uma linha mais recente, Halder et al. [Halder et al. 2022] propõem um modelo multitarefa baseado em *Transformers* para recomendar o próximo POI e predizer simultaneamente o tempo de espera.

Na classificação categórica, o TME [Xu et al. 2023] explora uma estrutura hierárquica de categorias para anotação semântica de POIs, enquanto o HMT-GRN [Lim et al. 2022] combina predição do próximo POI e de sua região geográfica em um modelo recorrente com grafos hierárquicos. Em conjunto, esses trabalhos indicam que o MTL é promissor em problemas de POIs, mas ainda há pouca exploração de arquiteturas que tratem conjuntamente a *classificação de categoria de POI* e a *predição do próximo POI* com representações espaciais contínuas e temporais explícitas. Essa lacuna motiva a investigação proposta neste trabalho.

3. Metodologia

Conforme discutido na Seção 2, o MTLNet [Silva et al. 2025] utiliza um único vetor $\mathbf{E}_{DGI} \in \mathbb{R}^{64}$ como entrada para ambas as tarefas, codificando informações espaciais e categóricas de forma conjunta por meio da topologia do grafo. Uma limitação dessa abordagem monolítica é que ela não incorpora coordenadas geográficas contínuas nem padrões temporais de visitação, fatores que, como demonstrado por trabalhos em codificação espacial [Wu et al. 2024] e temporal [Kazemi et al. 2019], carregam informações complementares às relações topológicas do grafo.

A decisão central de projeto deste trabalho é decompor a representação unimodal do DGI em três componentes independentes de 64 dimensões cada, treinados separadamente e integrados por concatenação: um *encoder* espacial contínuo, um *encoder* temporal (Time2Vec) e um *encoder* categórico hierárquico (HGI). A hipótese subjacente é que uma representação desacoplada, com *encoders* especializados treinados com funções de perda adequadas a cada dimensão da informação, produz representações mais expressivas do que uma representação unimodal que codifica todas as dimensões conjuntamente. Para avaliar essa hipótese e investigar o impacto da estratégia de codificação espacial, são comparados dois *encoders* com pressupostos arquiteturais distintos: SIREN e Sphere2Vec-M.

A metodologia proposta é ilustrada na Figura 1. A seguir são descritos o modelo MTLNet, o *baseline* com DGI, a justificativa e o funcionamento de cada *encoder* proposto, e a estratégia de integração dos *embeddings*.

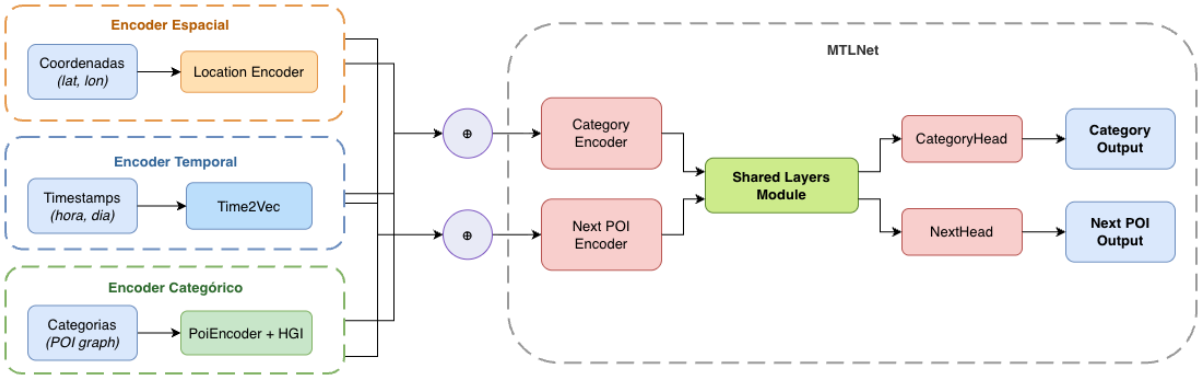

Figura 1. Arquitetura baseada no MTLNet [Silva et al. 2025]. Os *encoders* espacial, temporal e categórico são treinados de forma desacoplada e geram seus respectivos *embeddings*, que são integrados como entrada do modelo e processados pelas camadas compartilhadas do modelo multitarefa (*Shared Layers Module*) e pelas camadas específicas de cada tarefa, produzindo as saídas *Category Output* e *Next POI Output*.

O MTLNet [Silva et al. 2025] é uma rede multitarefa construída sobre compartilhamento rígido de parâmetros (*hard parameter sharing*) [Caruana 1997], que processa simultaneamente as tarefas de classificação de categoria de POI e predição do próximo POI. A arquitetura utiliza *encoders* específicos por tarefa implementados como MLPs com LayerNorm e Dropout, que projetam os *embeddings* de entrada em um espaço latente compartilhado de dimensão $d_{\text{shared}} = 256$. Em seguida, é aplicada a modulação FiLM (*Feature-wise Linear Modulation*) [Perez et al. 2018], condicionada por um *embedding* de identificador de tarefa $\mathbf{e}_t$, que gera parâmetros de escala γ_t e deslocamento β_t para modular as representações codificadas:

$$\text{FiLM}(\mathbf{h}_{\text{enc}}^{(t)} \,|\, \mathbf{e}_t) = \gamma_t \odot \mathbf{h}_{\text{enc}}^{(t)} + \beta_t \tag{1}$$

As representações moduladas são processadas por quatro blocos residuais compartilhados com LayerNorm, LeakyReLU e Dropout, permitindo ao modelo aprender uma representação comum para ambas as tarefas [Baxter 2000].

Na saída, cada tarefa possui uma *head* especializada. O módulo de classificação de categoria utiliza um conjunto de três MLPs paralelas com profundidades variadas (2, 3 e 4 camadas), cujas saídas são concatenadas e projetadas para as 7 classes de POI. A *head* de próximo POI emprega um *Transformer encoder* com 8 *heads* de atenção e 4 camadas, máscara causal para processamento autoregressivo, e *pooling* ponderado por pesos de atenção aprendidos sobre os *timesteps* da sequência de entrada.

O treinamento multitarefa utiliza o regularizador Nash-MTL [Navon et al. 2022], que formula o balanceamento dos gradientes como um jogo cooperativo de barganha de Nash entre K tarefas. Dados os gradientes de cada tarefa, o Nash-MTL busca a direção de atualização que maximiza o produto das utilidades de todas as tarefas, o que garante que a atualização seja benéfica para todas as tarefas simultaneamente, evitando a dominância de uma tarefa sobre a outra.

3.1. Baseline: MTLNet com DGI

O modelo *baseline* corresponde ao MTLNet original proposto em [Silva et al. 2025], cuja arquitetura interna é mantida inalterada neste trabalho. Nesse modelo, cada POI é representado por um único *embedding* monolítico $\mathbf{E}_{DGI} \in \mathbb{R}^{64}$, obtido via *Deep Graph Infomax* (DGI) [Veličković et al. 2019] sobre um grafo espacial construído por Triangulação de Delaunay e atributos categóricos dos nós, conforme descrito detalhadamente em [Silva et al. 2025]. Assim, o DGI codifica conjuntamente relações espaciais e categóricas em um único vetor, sem incorporar componentes temporais explícitas. Em contraste, a abordagem proposta substitui essa representação monolítica pela concatenação de três *encoders* especializados, totalizando 192 dimensões de entrada.

3.2. Preparação dos Dados

Para a tarefa de *classificação de categoria de POI*, cada POI é representado pelo *embedding* resultante da concatenação dos três componentes, $\mathbf{E}_{cat} = [\mathbf{E}_{HGI} \| \mathbf{E}_{loc} \| \mathbf{E}_{time}] \in \mathbb{R}^{192}$, formando pares $(\mathbf{E}_{cat}, c)$ onde c é a categoria real do POI. No caso do *baseline*, é utilizado diretamente $\mathbf{E}_{DGI} \in \mathbb{R}^{64}$.

Para a tarefa de *predição do próximo POI*, os *check-ins* são ordenados cronologicamente por usuário, e usuários com menos de cinco visitas são descartados. São extraídas janelas não sobrepostas de tamanho $L_h = 9$: os *check-ins* $(p_1, \ldots, p_9)$ formam o contexto, e p_{target} é o próximo POI cuja categoria deve ser predita. Cada *check-in* da janela é representado pelo *embedding* correspondente, resultando em uma entrada de $9 \times 192 = 1728$ dimensões para o ST-MTLNet e $9 \times 64 = 576$ dimensões para o *baseline*. Sequências mais curtas são preenchidas com *padding* de vetores nulos.

3.3. Encoder Espacial

O modelo original MTLNet codifica informações espaciais de forma implícita por meio da topologia do grafo de Delaunay, sem utilizar diretamente as coordenadas geográficas como sinal de entrada. No entanto, como discutido na Seção 2, codificadores espaciais contínuos têm demonstrado eficácia em diversas tarefas geoespaciais [Wu et al. 2024], embora sua aplicação em modelos multitarefa de POIs permaneça pouco explorada. Este trabalho seleciona dois *encoders* que representam paradigmas distintos de codificação espacial: o SIREN [Rußwurm et al. 2024], que modela funções contínuas por ativações senoidais com frequências controláveis, e o Sphere2Vec-M [Mai et al. 2023], que opera diretamente em coordenadas esféricas preservando propriedades de distância geodésica. Essa diversidade permite avaliar como diferentes pressupostos geométricos e arquiteturais impactam o desempenho em tarefas de POI.

Todos os *encoders* produzem $\mathbf{E}_{loc} \in \mathbb{R}^{64}$ e são treinados com a mesma função de perda contrastiva binária baseada na distância geográfica entre pares de coordenadas:

$$\mathcal{L}_{\text{loc}}(i,j) = -\left[y_{ij} \, \log \sigma\left(\frac{\text{sim}(z_i, z_j)}{\tau} \right) + (1 - y_{ij}) \log\left(1 - \sigma\left(\frac{\text{sim}(z_i, z_j)}{\tau} \right) \right) \right] \qquad (2)$$

na qual σ denota a função sigmoide, $\text{sim}(\cdot,\cdot)$ é a similaridade de cosseno, e $y_{ij} \in \{0,1\}$ indica se o par é positivo ($\Delta d_{ij} \leq 10$ km) ou negativo ($\Delta d_{ij} \geq 70$ km), com temperatura $\tau = 0{,}15$. A utilização da mesma função de perda para os dois *encoders* espaciais permite uma comparação controlada, isolando o efeito da arquitetura de codificação.

3.3.1. SIREN

O modelo SIREN (*Sinusoidal Representation Networks*) [Rußwurm et al. 2024] modela uma função contínua $f_\theta : \mathbb{R}^2 \to \mathbb{R}^{64}$ que mapeia diretamente coordenadas geográficas normalizadas em um *embedding* vetorial. Cada camada da rede utiliza ativações senoidais com frequências controláveis, permitindo representar sinais de alta frequência em dados espaciais. A saída é o *embedding* $\mathbf{E}_{loc} \in \mathbb{R}^{64}$.

3.3.2. Sphere2Vec-M

O Sphere2Vec-M [Mai et al. 2023] é um codificador multi-escala formulado para operar diretamente em coordenadas esféricas (λ, ϕ), gerando representações que preservem propriedades de distância na superfície da esfera. É utilizada a variante *sphereM*, na qual são empregadas $S = 16$ escalas multi-resolução com fatores geometricamente espaçados entre 10 km e 10.000 km. Termos de interação entre λ e ϕ são construídos mantendo uma das componentes no maior nível de escala, permitindo controlar a resolução angular de latitude e longitude de forma independente. O vetor resultante é projetado por uma MLP para produzir $\mathbf{E}_{loc} \in \mathbb{R}^{64}$.

3.4. Encoder Temporal

Padrões de mobilidade humana apresentam regularidades cíclicas, como horários de refeição e deslocamentos semanais, que carregam informações discriminativas sobre a natureza funcional dos POIs visitados [Sun et al. 2020]. No entanto, o modelo original MTLNet não incorpora representações temporais explícitas para capturar essas informações.

Para preencher essa lacuna, é utilizado o Time2Vec [Kazemi et al. 2019], uma representação temporal genérica e desacoplada capaz de capturar padrões periódicos e não periódicos sem necessidade de engenharia manual de *features*. A arquitetura combina termos lineares e periódicos para mapear os valores temporais de cada *check-in* (hora do dia e dia da semana, de forma acoplada) em um vetor de *embedding*. Para um *embedding* de dimensão D, a função é definida como:

$$\mathbf{f}(\tau)[i] = \begin{cases} \omega_0 \tau + \phi_0 & \text{se } i = 0 \\ \sin(\omega_i \tau + \phi_i) & \text{se } 1 \leq i \leq D-1 \end{cases} \tag{3}$$

na qual ω_i e ϕ_i são parâmetros treináveis (pesos e *bias*) que capturam a periodicidade dos padrões temporais de visitação. O termo linear ($i = 0$) captura tendências temporais globais, enquanto os termos senoidais modelam padrões cíclicos em diferentes frequências.

O modelo é treinado usando os valores temporais discretos (hora do dia e dia da semana) de forma acoplada para cada *check-in*, com a mesma formulação contrastiva binária empregada no encoder espacial, sendo denominada $\mathcal{L}_{tempo}$. A saída final é o *embedding* $\mathbf{E}_{time} \in \mathbb{R}^{64}$, que representa o *timestamp* de cada *check-in*.

3.5. Encoder Categórico

O DGI utilizado no modelo original codifica informações categóricas por meio de uma matriz *one-hot* de 7 classes processada pela estrutura do grafo, sem capturar explicitamente relações hierárquicas ou regionais entre categorias de POIs. Para enriquecer essa dimensão, a geração de *embeddings* categóricos é realizada em duas fases complementares. O POI Encoder

aprende representações vetoriais que capturam coocorrências entre categorias a partir de caminhadas aleatórias no grafo espacial, e o modelo HGI [Huang et al. 2023] incorpora informações hierárquicas em múltiplos níveis (POI, região e cidade), produzindo representações que refletem tanto o contexto local quanto o regional.

3.5.1. POI Encoder

O POI Encoder é treinado de forma independente para gerar o *embedding* de categoria $\mathbf{E}_{\text{POI\_categoria}} \in \mathbb{R}^{64}$. Os POIs são organizados em um grafo espacial construído por Triangulação de Delaunay sobre as coordenadas geográficas, com arestas ponderadas por decaimento logarítmico da distância Haversine e penalização para conexões entre condados (GEOIDs) distintos.

Sobre esse grafo são executadas caminhadas aleatórias (*random walks*) seguindo a metodologia do Node2Vec [Grover and Leskovec 2016]. Cada caminhada é convertida em uma sequência de categorias secundárias (*fclass*), resultando em coocorrências locais e globais entre categorias. O modelo aprende os *embeddings* utilizando a estratégia *skip-gram* com *negative sampling* [Church 2017]:

$$\mathcal{L}_{\text{SGNS}}(i,j) = -\log \sigma\big(\mathbf{e}_j^\top \mathbf{e}_i\big) - \sum_{k=1}^{K} \log \sigma\big(-\mathbf{e}_{n_k}^\top \mathbf{e}_i\big)\,,$$

na qual $\mathbf{e}_i$ é o *embedding* da classe alvo, $\mathbf{e}_j$ o *embedding* positivo de contexto, e $\mathbf{e}_{n_k}$ representam as amostras negativas. A implementação incorpora um termo de regularização hierárquica [Belkin and Niyogi 2003] entre categoria e *fclass*: $\mathcal{L}_{\text{hier}} = \sum_{(c,s)\in\mathcal{H}} \|\mathbf{e}_s - \mathbf{e}_c\|_2^2$, na qual $\mathcal{H}$ contém os pares (categoria, *fclass*). A perda total é $\mathcal{L} = \mathcal{L}_{\text{SGNS}} + \lambda_{\text{hier}}\,\mathcal{L}_{\text{hier}}$. Ao final, o *embedding* resultante é gerado por categoria e remapeado para cada POI.

3.5.2. Hierarchical Graph Infomax (HGI)

Após a geração dos *embeddings* $\mathbf{E}_{\text{POI\_categoria}}$, eles são utilizados como entrada para o modelo Hierarchical Graph Infomax (HGI) [Huang et al. 2023], que enriquece a representação incorporando informações geográficas regionais. O modelo otimiza a informação mútua em dois níveis hierárquicos (POI-região e região-cidade):

$$\mathcal{L} = \alpha\,\mathcal{L}_{\text{POI–região}} + (1-\alpha)\,\mathcal{L}_{\text{região–cidade}},$$

na qual $\alpha = 0{,}5$ controla o equilíbrio entre os níveis hierárquicos. Ambas as perdas utilizam amostragem negativa estrutural para reforçar a separabilidade entre regiões e entre POIs. A saída são os *embeddings* enriquecidos $\mathbf{E}_{\text{HGI}} \in \mathbb{R}^{64}$, que capturam relações categóricas tanto locais quanto regionais.

3.6. Integração dos Embeddings

Os três componentes de representação são treinados de forma independente e concatenados para formar a entrada de ambas as tarefas do MTLNet:

$$\mathbf{E}_{input} = [\mathbf{E}_{HGI} \| \mathbf{E}_{loc} \| \mathbf{E}_{time}] \in \mathbb{R}^{192} \tag{4}$$

A escolha da concatenação como estratégia de integração preserva a informação individual de cada componente sem impor interações prematuras entre as representações, delegando ao MTLNet a tarefa de aprender como combinar essas dimensões por meio de suas camadas compartilhadas e da modulação FiLM. A dimensionalidade de entrada do ST-MTLNet ($\mathbb{R}^{192}$) é superior à do *baseline* ($\mathbb{R}^{64}$), uma consequência natural da decomposição em três componentes especializados.

No entanto, a arquitetura do MTLNet projeta qualquer entrada para o mesmo espaço latente compartilhado de dimensão $d_{\text{shared}} = 256$ por meio dos *encoders* específicos por tarefa, de modo que a capacidade das camadas compartilhadas e das cabeças de tarefa permanece inalterada entre os modelos avaliados. Ainda assim, a diferença na dimensionalidade de entrada pode influenciar parte dos ganhos observados. Por isso, um controle experimental adicional igualando a dimensionalidade das representações permitiria validar de forma mais precisa se os ganhos ocorrem principalmente pela especialização semântica dos *encoders*, e não apenas pelo aumento da dimensionalidade de entrada.

São avaliadas duas variantes do modelo proposto, denominadas ST-MTLNet (*Spatial-Temporal MTLNet*), cada uma utilizando um *encoder* espacial diferente. O *baseline* MTLNet [Silva et al. 2025] utiliza somente o *embedding* monolítico $\mathbf{E}_{DGI} \in \mathbb{R}^{64}$ como entrada para ambas as tarefas. As variantes propostas substituem essa representação pela concatenação dos *embeddings* desacoplados, mantendo fixos os *encoders* temporal (Time2Vec) e categórico (HGI) e variando o *encoder* espacial entre SIREN e Sphere2Vec-M.

Todos os modelos compartilham os mesmos hiperparâmetros da arquitetura MTLNet, diferindo apenas nos *embeddings* de entrada. Essa configuração permite isolar o efeito da estratégia de representação sobre o desempenho, avaliando tanto o ganho da abordagem modular em relação ao *baseline* monolítico quanto o impacto da escolha do *encoder* espacial.

4. Resultados

4.1. Configuração Experimental

O desempenho dos modelos é avaliado por meio do F1-Score médio por categoria, reportado como média e desvio padrão em 5 *folds* com divisão estratificada, utilizando 80% dos dados para treino e 20% para validação, preservando a proporção entre categorias.

Tabela 1. Total de *check-ins*, POIs, e usuários de Flórida, Califórnia e Texas.

Estados	*Check-ins*	Pontos de Interesse	Usuários
Florida	990.518	65.009	20.301
California	2.535.573	148.314	36.106
Texas	3.355.419	135.570	37.522

Os experimentos foram conduzidos com o *dataset* Gowalla [Liu et al. 2014] nos estados da Flórida, Califórnia e Texas, cujas estatísticas são apresentadas na Tabela 1. A escolha desses estados se deve ao elevado volume de *check-ins* e à presença de padrões urbanos e distribuições espaciais distintas, o que permite avaliar o comportamento dos modelos em cenários com diferentes níveis de densidade e dispersão geográfica.

4.2. Classificação de Categoria de POI

Os resultados da tarefa de classificação de categoria de POI são apresentados na Tabela 2. Nos três estados avaliados, as duas variantes do ST-MTLNet superam o MTLNet original

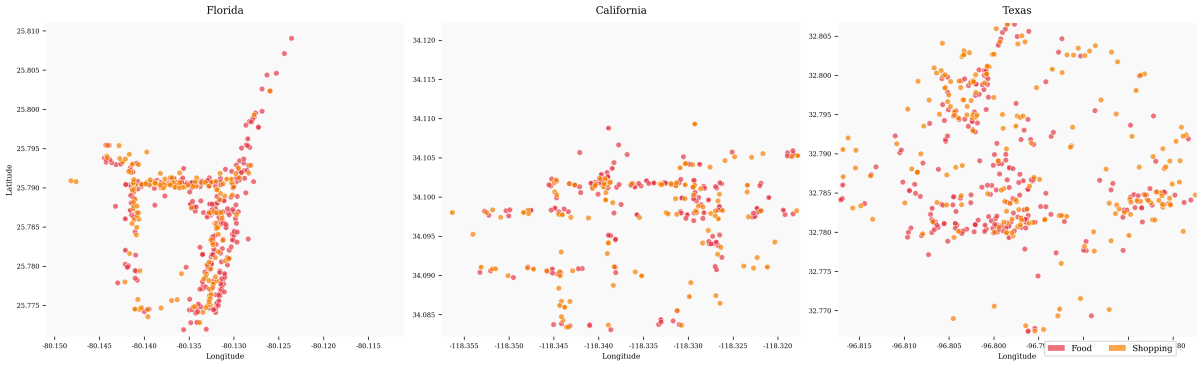

Figura 2. Distribuição espacial de POIs das categorias Food (vermelho) e Shopping (laranja) na sub-região mais densa de Flórida, Califórnia e Texas. As sub-regiões foram selecionadas por densidade em *grid*, com cerca de 100 POIs por região.

em todas as 21 combinações de categoria e estado, evidenciando que a substituição do *embedding* monolítico do DGI por representações desacopladas enriquece de forma consistente a caracterização dos POIs. Os ganhos são particularmente expressivos em categorias como *Nightlife* e *Travel*. Em *Nightlife*, por exemplo, o ST-MTLNet com Sphere2Vec-M atinge 62,44 $\pm$ 2,00 na Flórida, 60,78 $\pm$ 1,63 na Califórnia e 64,57 $\pm$ 1,03 no Texas, sempre muito acima do *baseline*. Em *Travel*, os melhores resultados chegam a 64,89 $\pm$ 1,20 na Flórida com SIREN, 63,59 $\pm$ 1,45 na Califórnia com SIREN e 64,73 $\pm$ 0,72 no Texas com Sphere2Vec-M.

As melhorias observadas em *Food* e *Shopping* reforçam que a localização geográfica carrega sinais discriminativos relevantes para a categoria semântica do POI. A Figura 2 mostra que essas categorias tendem a formar aglomerados densos nas sub-regiões analisadas, favorecendo a captura de padrões espaciais pelos *encoders* contínuos. Embora as duas variantes propostas apresentem desempenho próximo, observa-se um comportamento complementar: o SIREN obtém mais resultados de destaque na Flórida e na Califórnia, enquanto o Sphere2Vec-M se destaca com maior frequência no Texas. Em conjunto, esses resultados mostram que a abordagem modular supera de forma consistente o paradigma monolítico do DGI nessa tarefa.

4.3. Predição do Próximo POI

Os resultados da tarefa de predição do próximo POI são apresentados na Tabela 3. Diferentemente do que ocorre na classificação, essa tarefa apresenta um cenário mais heterogêneo, mas ainda favorável às variantes propostas: os modelos espaço-temporais superam o MTLNet original em 16 das 21 combinações avaliadas. Os ganhos mais expressivos ocorrem em *Food*, categoria em que ambas as variantes superam o *baseline* nos três estados. O melhor resultado é obtido pelo Sphere2Vec-M no Texas, com 48,67 $\pm$ 0,66, enquanto na Califórnia o mesmo modelo alcança 51,28 $\pm$ 0,57, superando com ampla margem o DGI. Melhorias consistentes também aparecem em *Shopping* e *Community*, sugerindo que a combinação entre informação espacial contínua e contexto temporal favorece a modelagem de transições recorrentes entre categorias.

O padrão de co-localização mostrado na Figura 2 ajuda a interpretar esse comportamento, pois a presença de agrupamentos densos de categorias como *Food* e *Shopping* favorece transições locais frequentes ao longo da trajetória do usuário. Ainda assim, o *baseline* mantém vantagem em alguns casos, especialmente em *Travel* na Flórida e na Califórnia, além de categorias específicas na Califórnia, como *Entertainment*, *Nightlife* e *Outdoors*.

Esse comportamento pode estar relacionado à própria natureza da categoria *Travel*, que

Tabela 2. F1-Score Médio (%) por modelo e estado para a tarefa de Classificação de Categoria de POI.

		MTLNet	**ST-MTLNet$_{SIREN}$**	**ST-MTLNet$_{Sphere2Vec-M}$**
Florida	Community	51,86 ± 0,73	**70,00 ± 0,81**	69,84 ± 0,98
	Entertainment	41,24 ± 1,83	**64,45 ± 1,82**	63,92 ± 1,61
	Food	55,47 ± 1,55	72,92 ± 0,59	**73,22 ± 0,66**
	Nightlife	32,59 ± 1,96	**62,60 ± 1,96**	62,44 ± 2,00
	Outdoors	47,71 ± 1,60	65,64 ± 1,73	**66,35 ± 1,51**
	Shopping	62,96 ± 0,62	**77,48 ± 0,36**	77,47 ± 0,72
	Travel	45,49 ± 1,20	**64,89 ± 1,20**	64,77 ± 1,54
California	Community	53,22 ± 0,37	**70,21 ± 0,17**	70,08 ± 0,44
	Entertainment	40,89 ± 1,32	**63,60 ± 0,83**	63,21 ± 0,76
	Food	60,68 ± 0,69	**75,78 ± 0,39**	75,74 ± 0,19
	Nightlife	26,81 ± 1,71	60,59 ± 1,35	**60,78 ± 1,63**
	Outdoors	51,59 ± 1,15	**67,94 ± 1,52**	67,37 ± 1,95
	Shopping	60,43 ± 1,01	76,84 ± 0,29	**77,00 ± 0,33**
	Travel	38,88 ± 1,32	**63,59 ± 1,45**	63,25 ± 0,99
Texas	Community	57,37 ± 0,69	73,45 ± 0,34	**76,20 ± 0,46**
	Entertainment	46,69 ± 0,46	64,09 ± 1,24	**67,73 ± 1,43**
	Food	54,31 ± 0,42	69,80 ± 0,80	**73,10 ± 0,17**
	Nightlife	34,44 ± 1,21	60,62 ± 1,32	**64,57 ± 1,03**
	Outdoors	45,99 ± 1,70	64,21 ± 0,42	**68,72 ± 0,70**
	Shopping	62,70 ± 0,29	76,40 ± 0,44	**79,67 ± 0,14**
	Travel	39,37 ± 1,29	57,53 ± 1,08	**64,73 ± 0,72**

tende a envolver deslocamentos mais esparsos entre regiões distantes. Nesses casos, a topologia do grafo utilizada pelo DGI pode ser mais eficiente para a preservação de relações entre POIs geograficamente distantes entre si. Por outro lado, *encoders* espaciais como SIREN e Sphere2Vec-M são baseados em coordenadas, sendo mais adequados para capturar padrões locais, mostrando que codificações contínuas de coordenadas e representações baseadas em grafo capturam aspectos complementares da mobilidade. Mesmo com essa limitação, o desempenho geral indica que as representações propostas são mais eficazes na maioria dos cenários de predição sequencial.

5. Conclusão e Trabalhos Futuros

Este trabalho investigou se a substituição do *embedding* monolítico do DGI por representações desacopladas e especializadas poderia produzir uma base mais adequada para as tarefas de classificação de categoria de POI e predição do próximo POI. Os resultados obtidos nos três estados avaliados indicam que sim. Ao incorporar um *encoder* espacial contínuo, um *encoder* temporal (Time2Vec) e um *encoder* categórico hierárquico (HGI), o ST-MTLNet supera de forma consistente o *baseline* baseado apenas em DGI.

Na tarefa de classificação de categoria de POI, as variantes propostas superam o MTLNet em todas as 21 combinações de categoria e estado, com ganhos médios de 20 a 24 pontos percentuais. Na tarefa de predição do próximo POI, o desempenho também é majoritariamente favorável ao ST-MTLNet, que vence em 16 das 21 combinações avaliadas, com destaque para categorias como *Food*. Em conjunto, esses resultados mostram que a decomposição da representação em componentes espaciais, temporais e categóricos permite capturar padrões

Tabela 3. F1-Score Médio (%) por modelo e estado para a tarefa de Predição de Próximo POI.

		MTLNet	**ST-MTLNet$_{SIREN}$**	**ST-MTLNet$_{Sphere2Vec-M}$**
Florida	Community	34,33 ± 1,33	**38,31 ± 0,90**	37,98 ± 1,22
	Entertainment	25,34 ± 1,07	**31,38 ± 1,72**	31,24 ± 2,41
	Food	27,12 ± 2,79	**41,91 ± 1,93**	41,65 ± 3,15
	Nightlife	21,33 ± 1,61	23,32 ± 2,74	**23,98 ± 1,94**
	Outdoors	**21,61 ± 0,99**	21,29 ± 1,59	21,59 ± 1,91
	Shopping	39,62 ± 3,40	44,00 ± 2,04	**44,66 ± 2,38**
	Travel	**64,47 ± 1,02**	45,00 ± 1,10	44,93 ± 1,11
California	Community	32,14 ± 0,89	**37,86 ± 0,20**	37,80 ± 1,11
	Entertainment	**19,38 ± 0,94**	17,28 ± 1,73	17,01 ± 1,39
	Food	29,26 ± 1,67	50,42 ± 1,15	**51,28 ± 0,57**
	Nightlife	**18,24 ± 0,53**	16,34 ± 1,95	15,10 ± 2,70
	Outdoors	**25,01 ± 0,81**	24,84 ± 1,70	24,07 ± 2,46
	Shopping	38,41 ± 2,64	**39,04 ± 1,72**	37,82 ± 1,31
	Travel	**46,05 ± 0,84**	36,94 ± 1,70	37,82 ± 1,04
Texas	Community	34,22 ± 0,43	35,99 ± 0,93	**36,97 ± 0,81**
	Entertainment	25,56 ± 1,56	**28,30 ± 1,37**	27,59 ± 1,64
	Food	29,56 ± 4,10	47,48 ± 2,42	**48,67 ± 0,66**
	Nightlife	25,46 ± 1,13	26,06 ± 1,36	**26,14 ± 2,66**
	Outdoors	23,02 ± 1,03	**24,62 ± 0,46**	23,04 ± 1,44
	Shopping	38,79 ± 4,12	**45,40 ± 1,30**	43,18 ± 2,19
	Travel	29,71 ± 1,32	**34,26 ± 0,90**	33,05 ± 1,51

que o DGI, por si só, não modela de forma suficiente.

A comparação entre SIREN e Sphere2Vec-M também mostra que não há um único *encoder* espacial universalmente superior. O SIREN apresenta melhor desempenho com maior frequência na Flórida e na Califórnia, enquanto o Sphere2Vec-M se destaca no Texas, sugerindo que a adequação do *encoder* depende da distribuição geográfica dos POIs em cada território. Assim, o principal ganho do trabalho não está apenas na escolha de uma arquitetura espacial específica, mas na própria adoção de representações desacopladas como alternativa ao paradigma monolítico do DGI.

Uma limitação relevante permanece na categoria *Travel* da tarefa de predição do próximo POI, na qual o MTLNet original ainda obtém os melhores resultados em parte dos cenários. Isso sugere que deslocamentos de longa distância e relações topológicas entre regiões ainda são melhor capturados pela estrutura de grafo utilizada pelo DGI. Além disso, como os três componentes propostos são utilizados em conjunto, este trabalho não isola a contribuição individual de cada *encoder*, o que restringe uma análise mais detalhada sobre o peso relativo de cada fonte de informação.

Outra limitação está relacionada ao uso exclusivo do Gowalla, uma base comumente utilizada na literatura, mas composta por registros de mobilidade coletados entre fevereiro de 2009 e outubro de 2010. Com isso, os resultados devem ser interpretados com cautela quanto à sua generalização para padrões atuais de mobilidade urbana.

Como trabalhos futuros, destacam-se a investigação de arquiteturas multitarefa mais flexíveis, a exploração de estratégias de fusão mais sofisticadas entre os *encoders* e a

combinação entre representações baseadas em grafo e codificações contínuas de coordenadas. Essas direções podem ampliar os ganhos observados e ajudar a reduzir as limitações ainda presentes em categorias como *Travel*.

Agradecimentos

Os autores gostariam de agradecer o apoio do Ministério da Ciência, Tecnologia e Inovações (MCTI), da Equipe Manna, da Fundação Araucária, da Softex, do CNPq (número do projeto 421548/2022-3), da Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) e da CAPES.

Referências

Baxter, J. (2000). A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198.

Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. volume 15, pages 1373–1396.

Caruana, R. (1997). Multitask learning. *Machine Learning*, 28(1):41–75.

Cho, E., Myers, S. A., and Leskovec, J. (2011). Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090.

Church, K. W. (2017). Word2vec. *Natural Language Engineering*, 23(1):155–162.

Feng, S., Cong, G., An, B., and Chee, Y. M. (2017). Poi2vec: Geographical latent representation for predicting future visitors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Grover, A. and Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 855–864.

Halder, S., Lim, K. H., Chan, J., and Zhang, X. (2022). Poi recommendation with queuing time and user interest awareness. *Data Mining and Knowledge Discovery*, 36:2379–2409.

Huang, W., Zhang, D., Mai, G., Guo, X., and Cui, L. (2023). Learning urban region representations with pois and hierarchical graph infomax. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196:134–145.

Jure, L. (2014). Snap datasets: Stanford large network dataset collection. *Retrieved December 2021 from http://snap. stanford. edu/data*.

Kazemi, S. M., Goel, R., Eghbali, S., Ramanan, J., Sahota, J., Thakur, S., Wu, S., Smyth, C., Poupart, P., and Brubaker, M. (2019). Time2vec: Learning a vector representation of time. *arXiv preprint arXiv:1907.05321*.

Liao, D., Liu, W., Zhong, Y., Li, J., and Wang, G. (2018). Predicting activity and location with multi-task context aware recurrent neural network. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 3435–3441. International Joint Conferences on Artificial Intelligence Organization.

Lim, N., Hooi, B., Ng, S.-K., Goh, Y. L., Weng, R., and Tan, R. (2022). Hierarchical multi-task graph recurrent network for next poi recommendation. In *SIGIR '22: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1133–1143. ACM.

Liu, Y., Wei, W., Sun, A., and Miao, C. (2014). Exploiting geographical neighborhood characteristics for location recommendation. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, page 739–748, New York, NY, USA. Association for Computing Machinery.

Mai, G., Xuan, Y., Zuo, W., He, Y., Song, J., Ermon, S., Janowicz, K., and Lao, N. (2023). Sphere2vec: A general-purpose location representation learning over a spherical surface for large-scale geospatial predictions.

Navon, A., Shamsian, A., Achituve, I., Maron, H., Kawaguchi, K., Chechik, G., and Fetaya, E. (2022). Multi-task learning as a bargaining game. *arXiv preprint arXiv:2202.01017*.

Perez, E., Strub, F., de Vries, H., Dumoulin, V., and Courville, A. C. (2018). FiLM: Visual reasoning with a general conditioning layer. In *Proc. AAAI Conf. Artificial Intelligence*, pages 3942–3951.

Rahmani, H. A., Aliannejadi, M., Mirzaei Zadeh, R., Baratchi, M., Afsharchi, M., and Crestani, F. (2019). Category-aware location embedding for point-of-interest recommendation. In *Proceedings of the 2019 ACM SIGIR international conference on theory of information retrieval*, pages 173–176.

Rußwurm, M., Klemmer, K., Rolf, E., Zbinden, R., and Tuia, D. (2024). Geographic location encoding with spherical harmonics and sinusoidal representation networks.

Silva, V. H. O., Almeida, I. F., Paiva, T. S., Santos, G. B., Silva, F. A., and Sousa, F. T. (2025). An investigation into multi-task learning for point-of-interest category classification and next-poi prediction. In *Proceedings of the Brazilian Conference on Intelligent Systems (CBIC)*. Submetido.

Sitzmann, V., Martel, J., Bergman, A., Lindell, D., and Wetzstein, G. (2020). Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems*, volume 33, pages 7462–7473.

Sun, K., Qian, T., Chen, T., Liang, Y., Nguyen, Q. V. H., and Yin, H. (2020). Where to go next: Modeling long-and short-term user preferences for point-of-interest recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 214–221.

Sun, Y. (2024). Transtarec: Time-adaptive translating embedding model for next poi recommendation. In *2024 5th International Conference on Computer Engineering and Application (ICCEA)*, pages 647–651. IEEE.

Veličković, P., Fedus, W., Hamilton, W. L., Liò, P., Bengio, Y., and Hjelm, R. D. (2019). Deep graph infomax. In *International Conference on Learning Representations (ICLR)*.

Wu, N., Cao, Q., Wang, Z., Liu, Z., Qi, Y., Zhang, J., Ni, J., Yao, X., Ma, H., Mu, L., et al. (2024). Torchspatial: A location encoding framework and benchmark for spatial representation learning. *Advances in Neural Information Processing Systems*, 37:81437–81460.

Xia, B., Bai, Y., Yin, J., Li, Q., and Xu, L. (2020). Mtpr: A multi-task learning based poi recommendation considering temporal check-ins and geographical locations. *Applied Sciences*, 10(19):6664.

Xu, R., Chen, M., Gong, Y., Liu, Y., Yu, X., and Nie, L. (2023). Tme: Tree-guided multi-task embedding learning towards semantic venue annotation. *ACM Trans. Inf. Syst.*, 41(4):112.