AddressBind: Cross-modal Alignment of Addresses and Geocodes

Govind
gvindmg@amazon.com
Amazon
Bangalore, India

Sayan Putatunda
putatund@amazon.com
Amazon
Bangalore, India

Saurabh Sohoney
sohoneys@amazon.com
Amazon
Hyderabad, India

Abstract

Mapping addresses to geolocations accurately is a challenging and important problem, with many real-world applications such as delivery logistics, map building and path finding. High quality embedding of geospatial data (e.g., addresses, geocodes) which is grounded in real world play an important role in success of modeling tasks such as geocoding and address resolution/matching. Existing state-of-the-art (SOTA) approaches [9] have proposed to transform the address embedding space to mimic real world proximity via a triplet loss, but requires triplet engineering which is error prone and difficult to scale. In this work, we propose to embed addresses and geocodes data in the same embedding space to enable late fusion of cross-modal semantics and remove dependency on triplet creation. Our proposed model outperforms SOTA baselines (including Multilingual-E5-Large-Instruct [32], a top model on MTEB leaderboard) by improving geolocation accuracy and geocode outliers across geographies with diverse writing standards. We also observe significant gains in address embeddings quality intrinsically and the approach supports to jointly align more geospatial modalities.

CCS Concepts

• Computing methodologies → Natural language processing; **• Information systems → Document representation**; **Multimedia and multimodal retrieval**; **Location based services**; **Geographic information systems**; **Language models**; **Nearest-neighbor search**.

Keywords

Geospatial semantics, Cross-modal representation learning, Address-geocode alignment, Spatial proximity modeling

ACM Reference Format:
Govind, Sayan Putatunda, and Saurabh Sohoney. 2025. AddressBind: Cross-modal Alignment of Addresses and Geocodes. In *The 33rd ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '25), November 3–6, 2025, Minneapolis, MN, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3748636.3764598

1 Introduction and Motivation

Learning accurate geolocation of addresses is important for various business-to-consumer services such as logistics, ride-hailing, and emergency response operations. The process of determining the geocode (i.e., latitude and longitude) of addresses is commonly known as geocoding and usually accomplished by matching the query address text to a known addresses reference set. Therefore, it is of utmost importance to build a comprehensive understanding of the address text in order to accurately match it to a known reference set. Address text in regions where a standard structure is followed (such as North America and Europe) is largely driven by street number and street name but can be non-trivial to understand by machine learning (ML) models due to the presence of ambiguous numbers and instructions such as *X 3 Lakes Rd, Three lakes rd XXXX left side of street, 60010, Barrington, IL, US*[1]. On the other hand, many geographies have loosely structured addresses or a variety of local standards which lead to spelling variations, use of local languages, synonyms, abbreviations, etc. For example, the address *Sonai nadi ka pul, Sonai Baidpura Etawah, 206002, Etawah, UP, IN* uses romanized Hindi and does not contain any fine-grained details beyond the locality and landmark (i.e., *river bridge in Sonai village*). Popular commercial map services encounter challenges when processing addresses with these characteristics [27] and fail to correctly locate the above discussed addresses. Understanding addresses accurately is not just a technical challenge but has significant real-world implications. Inaccuracy in geocodes could lead to egregious planning errors, such as not being able to respond to emergencies promptly or creating poor experiences when directed to incorrect locations. Furthermore, high-quality address and geocode embeddings enable critical applications beyond geocoding and vector based candidates retrieval for address matching systems. Embedding-based clustering of co-located addresses provides analytical insights that can enable data-driven urban planning decisions. Also, cross-modal matching can help identify potentially erroneous address-geocode pairs that could affect aforementioned delivery services.

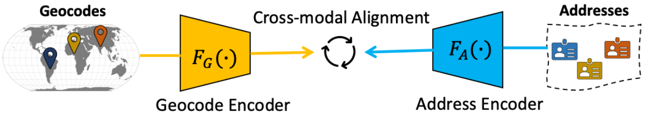


Figure 1: Jointly aligning geocodes and address texts

We observe that the standard sentence embedding approaches in natural language processing (NLP) literature, which are trained with language modeling objective [17] and/or a contrastive objective on sentence level [6, 25], fail to provide quality embeddings for addresses. We also note that SOTA instruction-tunable embedding models such as Multilingual-E5-Large-Instruct [32] which is in top-5 high performant models on MTEB leaderboard [1, 5], also struggle to provide quality embeddings for addresses (cf. Section

[1] Finer address details/numbers are masked (X) to preserve privacy.

SIGSPATIAL '25, Minneapolis, MN, USA

ACM ISBN 979-8-4007-2086-4/2025/11
https://doi.org/10.1145/3748636.3764598

4.1). The key reason is that address data is not organized as paragraphs/documents and thus, lacks context by nature. Instead, addresses derive their semantics primarily from the spatial relationships. One of the recent work on address representation learning [9] (referred as *RoBERTa-Triplet-H3* hereafter) tries to bridge this gap by transforming the embedding space to mimic real world proximity among addresses with the help of triplet loss. Here, the triplets are engineered using the H3 spatial index[2] and subjected to the constraint that the anchor address should be closer to the positive address in real world than to the negative address. This approach requires triplets engineering with some geography specific domain knowledge and can be difficult to scale.

In this paper, we propose AddressBind to bind address and geocode modalities by jointly embedding them in the same high dimensional space allowing richer transfer of proximity semantics to address embeddings. It eliminates the need of triplet engineering and leverages cross-modal learning to align representations via a self-supervised matching objective as illustrated in Figure 1. This enables the late fusion of knowledge across the two modalities. As a result, address representations will learn to associate to geocodes based on the various grounding tokens present in addresses. At the same time, geocode representations will learn to associate a given geocode with local addresses' characteristics in its vicinity. Throughout this paper we use the terms 'location' and 'geocode' interchangeably, as well as the terms 'embedding' and 'representation'. In summary, the key contributions of this paper are:

- To the best of our knowledge, we are the first to align the continuous encoding of addresses and geocodes in the same space to learn SOTA representations. It facilitates fusion of geospatial proximity into address embeddings and at the same time distills knowledge of local addresses in geocode embeddings.
- Our approach learns a continuous representation for addresses and geocodes allowing the model to interpolate to locations having no past deliveries. It eliminates the need for manual triplet engineering and is naturally extensible to include more geospatial modalities.
- We demonstrate the impact of our model on real-world applications across datasets that cover different writing styles and standards. This work focuses on tasks of address geocoding and anomalous geocode detection. However, learnt address and geocode embeddings can cater to other applications in related domain such as address correction, matching and entity recognition.

2 Related Work

Address Geocoding and Geospatial ML Address geocoding problem is recently gaining interest in both academic and industry research communities. Some of the recent works include [18], where Address geocoding is framed as pairwise matching problem using graph based active learning. Further, in [9], authors propose a dynamic neighbourhood level geocoding solution for cold-start hard-to-resolve addresses. In [21], a graph neural network based place representation learning solution is proposed for warm-start addresses. Qian et al. [22] experiment with a seq2seq geocoding model to directly predict geohash string for Chinese addresses. In [16], authors introduce GeoAttn model, which focuses on geolocation signals in the text and attends to the relevant Point-of-Interests (POIs) for location prediction. Srivastava et al. [28] propose to learn geospatial spread of terms in address text from delivery history data and predict geolocation based on their overlap. In computer vision domain, there are multiple recent works on learning from geospatial data such as geolocating wildlife images [19, 31] by embedding images with locations. In [12], authors propose to learn multi-purpose generic embeddings of world wide satellite imagery using a contrastive objective. Clark et al. [4] propose a discrete grid based classification approach to geolocate images. In [13], authors propose GeoChat, a large vision-language model to perform interactive remote sensing over satellite imagery. Learning robust embeddings via an auto-regressive denoising language modelling objective is also relevant for unstructured noisy address text [8]. In [26, 29], authors propose efficient way to model high frequency functions in low dimensional domains via random fourier features and sinusoidal representation networks. It is important to note that our work targets the fundamental problem of learning high-quality address embeddings, which enables multiple downstream applications such as geocoding and robust vector-based candidates retrieval for address matching and validation. This distinguishes our approach from methods that focus solely on specific tasks like building level address matching or geocoding without learning reusable embeddings. Such task-specific approaches, while effective for their intended purpose, can not be directly compared to our method as they optimize directly for downstream tasks rather than creating versatile representations that can support multiple geospatial applications.

Multi-modal Representation Learning In [23], authors introduce the CLIP model, which jointly trains an image encoder and a text encoder in a contrastive fashion to predict the correct pairings of (image, text) in training examples and learns transferable knowledge representations benefiting across multiple cross-modal tasks. Further in [14, 15], a vision-language pretraining framework is proposed based on multimodal mixture of encoder-decoder which employs a captioner to generate synthetic captions for web images, and a filter to remove noisy captions. GeoCLIP [31] take inspiration from CLIP to align images and geolocations in the same embedding space. Girdhar et al. [7] introduce ImageBind that aims to learn a joint embedding across six different modalities - images, text, audio, depth, thermal, and inertial measurement unit (IMU) data by using InfoNCE loss and observe emergent capabilities on unseen modalities pairs. Similarly, Zhu et al. [34] propose LanguageBind to align different modalities by keep language as the central modality to achieve cross-modal semantic alignment. In [8], authors propose robust representation learning of noisy text via an auto-regressive language denoising task. To the best of our knowledge, ours is the first work that proposes to align the continuous encoding of geocodes and addresses in the same representation space for efficient address representation learning.

3 Proposed Model

With AddressBind, we aim to encode addresses and geocodes to a continuous higher dimensional space that 1) preserves proximity and hierarchical nature inherent to addresses, 2) fuses knowledge across

[2] H3 Spatial Index https://h3geo.org/

modalities. We train AddressBind with a cross-modal matching objective, the address encoder F_A is trained to associate an address a with a geocode g based on the various grounding cues present in addresses. At the same time, the location encoder F_G is trained to associate a given location with local vicinity specific address text characteristics.

3.1 Geocode Encoder

Neural models such as multi-layer perceptron (MLP) have been observed to have spectral bias when learning high frequency function in low dimensional domains such as coordinates data (e.g., geocodes, viewpoint coordinates in 3D scene reconstruction) [29]. In other words, what it means in reference to our work is that *address text can greatly vary with respect to a little change in geocode* making it a high frequency function to learn. Multiple recent works [20, 31, 33] have reported experimentally that a sinusoidal mapping of input coordinates enables MLP networks to learn higher frequency content, which is considered as a special case of Fourier features [24]. Figure 2 depicts the key components of the geocode encoder where inductive biases control spatial smoothness and hierarchical nature, allowing the model to interpolate to areas where no address-location pairs are present in the training data and preserve hierarchical relationship.

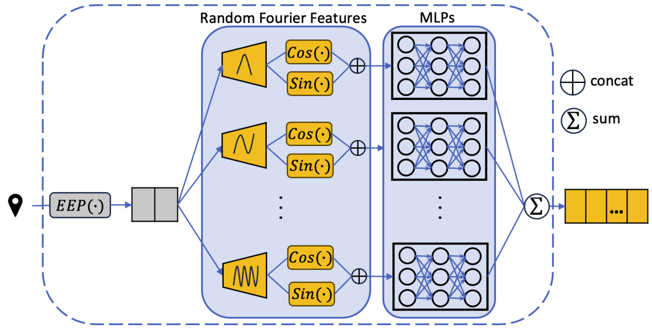


Figure 2: Encode geocode via Random Fourier Features

Random Fourier Features Equation 1 formulates the random fourier features (RFF) projection of a given geocode g from $\mathbb{R}^2$ into a higher dimensional space $\mathbb{R}^H; H \gg 2$ via a projection matrix J^σ sampled from the normal distribution $N(0, \sigma^2)$. The transformed geocode vector is fed to cosine and sine functions and concatenated. The value of σ controls the sensitivity of RFF projection towards fine or coarse-grained changes in geocodes (i.e., larger the value of σ, higher the sensitivity toward input changes). It should be noted that J^σ remains fixed throughout the training. Equal earth projection (EEP) [2] is performed over geocodes before input to RFF module, to reduce the impact of distortions in standard geographic coordinate system. Section 5.3 provides illustrations on the influence of σ on RFF sensitivity (cf. Fig. 9) and impact of not using RFF in encoder (cf. Fig. 8).

$$RFF(g, \sigma) = [cos(2\pi J^\sigma g), sin(2\pi J^\sigma g)]^T \quad (1)$$

Hierarchical Amalgamation An address text is generally influenced by multiple geographical entities of varying granularities. Therefore, it is a natural choice to have multiple RFFs (i.e. the hyperparameter Γ) with varying frequencies to obtain representations spanning from coarser to finer levels, which act as an useful inductive bias in the architecture. The frequency value σ_i for each of the RFF module is chosen using an exponential assignment strategy as suggested in [29, 31]. Subsequently, each of the RFF modules' projection are passed though MLPs independently. Finally, outputs from MLPs are aggregated via element-wise addition to produce the final geocode embedding. Equation 2 formally defines the location encoder F_G for a geocode g.

$$F_G(g) = \sum_{i=1}^{\Gamma} MLP_i(RFF(g, \sigma_i)) \quad (2)$$

3.2 Address Encoder

As an address encoder F_A, we utilize sentence-transformers [25] *all-MiniLM-L6-H384*[3], which is trained using contrastive learning on a very large sentence level dataset of 1B sentence pairs. The pretraining objective is: given a sentence from the pair, the model should predict which out of a set of randomly sampled other sentences, was actually paired with it in the dataset. We finetune the complete model weights during our cross-modal alignment training.

3.3 Contrastive Training

We leverage historical delivery data to train the proposed model using a self-supervised contrastive methodology (similar to CLIP [23, 31]) without being limited by manual data curation efforts. We aggregate historical delivery information to create (address, geocode) training pairs. Given that the geocode encoder can encode an arbitrary geocode and contrastive learning approaches benefit from large number of negatives in batch, we append $\mathbb{P}$ random geocode negative examples to every batch of size $\mathbb{B}$ similar to a general practice in literature [31]. Given a pair of address a_i and geocode g_i, the model is trained with InfoNCE [30] loss L_{a_i,g_i} to align address and geocode embeddings as formulated in Equation 3.

$$-log \frac{exp(F_A(a_i)^T F_G(g_i)/\tau)}{\sum_{j=1}^{\mathbb{B}} exp(F_A(a_i)^T F_G(g_j)/\tau) + \sum_{j=1}^{\mathbb{P}} exp(F_A(a_i)^T F_G(g_j)/\tau)} \quad (3)$$

Here, τ is a scalar temperature which controls smoothness of the softmax distribution (higher $\implies$ smoother). We note in experiments that adding extra geocode negative examples significantly improve the embeddings quality (cf. Section 4.1). We randomly sample geocode negatives on postal code group level and consider exploring further hard negative mining techniques as a future work. The described loss function can accommodate more geospatial modalities just by simply adding the corresponding paired data.

4 Experimental Analysis

We evaluate the learnt embeddings intrinsically, and on address geocoding and anomalous geocode detection tasks.

Datasets Details We experiment with addresses from structured (STR) as well as unstructured (USTR) geographical regions, and utilize historical delivery data for addresses to generate (address, geocode) pairs at large scale. Our model training leverages a large-scale dataset comprising tens of millions of unique addresses paired

[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Table 1: Confusion matrix showing 4 synthetic examples of (address, geocode) pairs for cross-modal alignment training. Using public landmarks for illustration, the matrix shows two types of negatives: in-batch (first 4 geocodes, where each address has one positive and three negatives) and extra negatives (last 2 geocodes) sample randomly from different locations

Address	**In-batch Geocodes** $\mathbb{B} = 4$				**Extra Negatives** $\mathbb{P} = 2$	
	40.689°N 74.045°W	*40.690°N 74.045°W*	*27.175°N 78.042°E*	*27.163°N 78.036°E*	*40.783°N 73.966°W*	*27.158°N 78.021°E*
Statue of Liberty, Liberty Island, New York, NY 10004, USA	+	−	−	−	−	−
Statue of Liberty Viewpoint, Jersey City, NJ 07305, USA	−	+	−	−	−	−
Taj Mahal Monument, Agra, UP 282001, India	−	−	+	−	−	−
Taj Mahal Restaurant, Sadar Bazaar, Agra, UP 282001, India	−	−	−	+	−	−

with their corresponding geocodes, derived from historical delivery data. For evaluation, few weeks of out-of-time network wide shipments (in hundreds of thousand) against addresses with no prior delivery are considered, and model predicted locations are compared against the observed delivery locations. As illustrated in Table 1, the training data consist of (address, geocode) pairs along with their positive or negative labels. Positive pairs are constructed from observed historical delivery locations, while negative pairs are constructed via in-batch negatives strategy. We augment each training batch of size $\mathbb{B}$ with extra $\mathbb{P}$ random geocode negatives to enhance the contrastive learning process. This approach enables the model to learn robust address-location relationships without requiring any manual data curation.

Model Configurations and Baselines For a thorough comparative analysis, we setup multiple baselines for address representation learning, models used in our experimentation can be grouped into the following three categories.

Sentence Embedding SOTA Models: We utilize sentence transformers [25] based *ST-MiniLM-L6-H384*, which is a general purpose sentence embedding model pretrained on a very large dataset of 1B sentence pairs. We also baseline against *Multilingual-E5-Large-Instruct* [32] model, which is a SOTA instruction-tuned sentence embedding model with strong performance on MTEB embeddings benchmark [5] and on par with other LLM based embedding models. We systematically tuned the *Multilingual-E5-Large-Instruct* instruction prompt for optimal address embedding performance, selecting:

> *Instruction: Given a query address, retrieve relevant addresses that are nearest to it in real world.*
> *Query: {query_address_text}*

ST-MiniLM-L6-H384 is 6 layers transformer based models with 384 embedding size, while *Multilingual-E5-Large-Instruct* is xlm-roberta-large based model with 1024 embedding size.

Address Embedding SOTA Models: We setup *RoBERTa-Address*, which is a RoBERTa [17] based model pretrained on addresses dataset with MLM objective. Further, we have another SOTA baseline *RoBERTa-Triplet-H3* [9] specifically proposed for learning address representation via H3 grids based triplets. To benchmark the performance without RFF projections, we have *MLP-Loc-Enc* model where the geocode input is directly passed to MLP module. *RoBERTa-Address* and *RoBERTa-Triplet-H3* are 6 layers RoBERTa based models with embedding size 768, while *MLP-Loc-Enc* is initialized from *ST-MiniLM-L6-H384*.

AddressBind Models: We have four variants of our AddressBind model. *AddressBind-4RFF* and *AddressBind-6RFF* are trained with the proposed geocode encoder having Γ values 4 and 6 respectively. Further, *AddressBind-6RFF-Noise* and *AddressBind-6RFF-4xNoise* models are trained with additionally supplied random geocode negatives of the size of batch and 4 times the size of batch respectively, as discussed in Section 3.3. AddressBind variants are 6 layers models with 384 embedding size initialized from *ST-MiniLM-L6-H384*.

Training and Implementation Details We use Adam optimizer [11] with learning rate of $5 * 10^{-5}$ and betas $(0.9, 0.999)$. We employ a linear learning rate schedular with step size of 1 epoch, start factor of 1.0 and end factor of 0.75. AddressBind models are trained with a batch size of 1024 until validation loss convergence or a maximum of 50 epochs. The batches are sampled from postal code group stratas to draw relevant in-batch negatives. The large batch size is particularly beneficial for contrastive learning as it provides more negative examples within each batch, improving the model's ability to distinguish between similar and dissimilar address-geocode pairs. In the following subsections, we report relative performance numbers w.r.t. the baseline wherever applicable in favor of business confidentiality without compromising on the experimentation rigor.

4.1 Embeddings Quality

We intrinsically evaluate the quality of embeddings learnt by different models w.r.t. the encoded real world proximity. We generate the proximity test sets for each geography by sampling 50K address pairs where both the addresses lie within a maximum of 50 m from each other. To avoid any data leakage or unfair advantage to any model, we only consider out of training addresses to generate these pairs. Given one of the address in pair, the task is to retrieve the other address by doing a K-Nearest Neighbour look up on embeddings. We facilitate this by using an approximate nearest neighbour tool

Table 2: Performance on address retrieval task to measure proximity semantics captured in embeddings. Results are shown for both structured (STR) and unstructured (USTR) geographical regions. Metrics with ↑ indicate higher is better.

	Model	hitrate@5 ↑		hitrate@10 ↑		hitrate@20 ↑		MRR@5 ↑		MRR@10 ↑		MRR@20 ↑	
		STR	USTR	STR	USTR	STR	USTR	STR	USTR	STR	USTR	STR	USTR
Baselines	ST-MiniLM-L6-v2	x	x	x	x	x	x	x	x	x	x	x	x
	Multilingual-E5-large-Instruct	1.109x	0.984x	1.151x	1.012x	1.202x	1.014x	1.08x	0.935x	1.093x	0.943x	1.104x	0.952x
	RoBERTa-Address	0.941x	1.098x	0.963x	1.205x	0.986x	1.320x	0.923x	1.032x	0.929x	1.059x	0.931x	1.114x
	RoBERTa-Triplet-H3	1.550x	3.000x	1.743x	3.247x	1.922x	3.454x	1.423x	2.581x	1.486x	2.676x	1.517x	2.800x
	MLP-Loc-Enc	0.986x	1.294x	1.098x	1.521x	1.233x	1.835x	0.923x	1.129x	0.957x	1.206x	0.979x	1.286x
Ours	**Final**:AddressBind-6RFF-4xNoise	**1.878x**	**3.686x**	**2.122x**	**4.205x**	**2.222x**	**4.536x**	**1.631x**	**3.097x**	**1.721x**	**3.265x**	**1.745x**	**3.429x**
	Base: w/ 4RFF	1.635x	3.137x	1.872x	3.575x	2.044x	3.979x	1.454x	2.677x	1.529x	2.853x	1.566x	3.000x
	w/ 6RFF	1.761x	3.314x	2.007x	3.781x	2.150x	4.196x	1.562x	2.774x	1.643x	2.941x	1.676x	3.114x
	w/ 6RFF, Noise	1.820x	3.451x	2.061x	3.932x	2.175x	4.340x	1.592x	2.871x	1.679x	3.059x	1.703x	3.229x

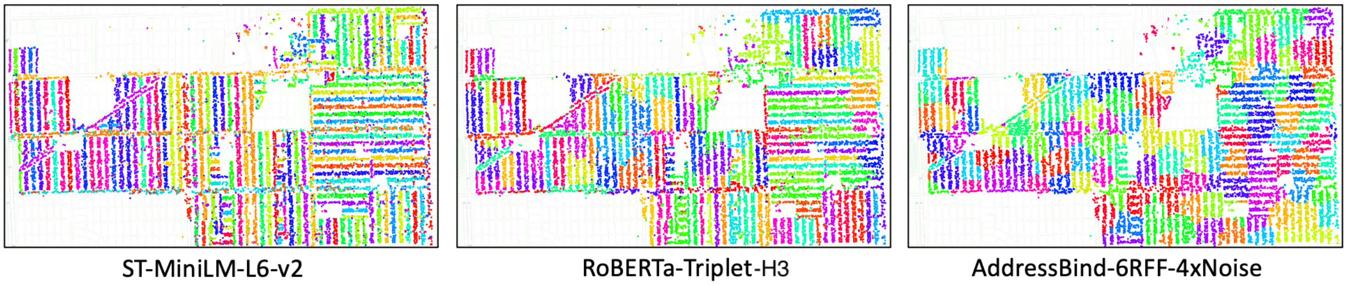


Figure 3: Clustering of addresses using embeddings from various models and visualization in geocode domain with colors depicting the discovered clusters. Silhouette scores are -0.32, -0.05 and 0.14 respectively for the depicted three models. (Note: Background maps are intentionally morphed to preserve privacy)

Annoy[4]. We utilize hit rate (hitrate@K) and mean reciprocal rank (MRR@K) as standard metrics [3] to measure the retrieval quality. The hit rate metric measures if we are able to retrieve a close proximity address for the query address in top-K list, whereas MRR further measures how high is the rank when there is a hit in top-K.

Table 2 reports relative values of hitrate@K and MRR@K w.r.t. the baseline for different K. We observe that AddressBind based models have superior performance than baselines, where *AddressBind-6RFF-4xNoise* performs the best among AddressBind variants, improving hitrate@5 over *RoBERTa-Triplet-H3* by >15% for both geographies. We observe a similar pattern in MRR metrics which implies that AddressBind models are not just capable of retrieving close proximity addresses successfully but also rank them better in the order. It is also worth noting that *MLP-Loc-Enc* model performs poorer than any of the RFF based models. Further, *Multilingual-E5-large-Instruct* performs poorer than *AddressBind* and *RoBERTa-Triplet-H3* models despite being a much larger model with SOTA performance on general purpose sentence embedding benchmarks, which highlights the need for special treatment for address domain. These empirical findings validate our hypothesis that AddressBind captures the real-world proximity better among addresses without relying on any triplet engineering.

Clustering Analysis We also perform a qualitative analysis by clustering (using K-means with K=100) addresses based on their embeddings and visualizing them through their geocodes (refer to Figure 3, a sample of 20K addresses). Here, the models which capture geospatial proximity more precisely, will result in smoother clusters by enabling the grouping of addresses together that are geographically closer. We observe that embeddings generated using *ST-MiniLM-L6-v2* produce clusters, which are only driven by street names in addresses. *RoBERTa-Triplet-H3* produces improved clusters due to slightly better encoding of proximity, but is still heavily driven by street names. *AddressBind-6RFF-4xNoise* produced clusters can be seen influenced by a good mix of real world proximity and address text, improving over the previous models. This is also visible in the Silhouette scores, which are -0.32, -0.05, and 0.14 for these models, respectively. The observed geospatial proximity semantics are beneficial for multiple geospatial tasks.

4.2 Address Geocode Learning

As an extrinsic evaluation, we compare models on the task of address geocoding. The geocode for a query address is predicted as an aggregated point over geocodes of retrieved matching addresses from the reference set, similar to a common practice in related literature such as image/address geolocation learning [9, 10], which allows for direct assessment of embedding quality without introducing additional model complexity. We utilize approximate nearest neighbour similarity search in our experimental analysis. Table 3 presents experimental results via various geocoding metrics on shipments for the test period. Acc@Y denotes accuracy of predicted geocodes falling within Y meters of actual geocodes. DR@Z (Defect Rate) measures geocoding outliers when the prediction falls beyond a certain

[4] Annoy Approximate Nearest Neighbours https://github.com/spotify/annoy

Table 3: Address geocoding performance comparison on structured (STR) and unstructured (USTR) geographical regions[5]. Metrics with ↑ indicate higher is better, while ↓ indicate lower is better.

	Model	Acc@Y ↑		DR@Z ↓		p25 ↓		p50 ↓		p95 ↓	
		STR	USTR	STR	USTR	STR	USTR	STR	USTR	STR	USTR
Baselines	ST-MiniLM-L6-v2	x	x	x	x	x	x	x	x	x	x
	Multilingual-E5-large-Instruct	1.01x	1.014x	0.853x	0.889x	0.921x	0.99x	0.983x	0.938x	0.369x	0.473x
	RoBERTa-Address	0.904x	1.015x	0.801x	0.630x	1.214x	1.095x	1.282x	0.838x	0.220x	0.204x
	RoBERTa-Triplet-H3	1.184x	1.634x	0.373x	0.427x	0.857x	0.476x	0.769x	0.283x	0.085x	0.153x
	MLP-Loc-Enc	0.969x	1.084x	0.625x	0.539x	1.286x	1.048x	1.128x	0.717x	0.140x	0.147x
Ours	**Final**: AddressBind-6RFF-4xNoise	**1.276x**	**1.704x**	0.313x	**0.371x**	**0.857x**	**0.476x**	**0.667x**	**0.253x**	**0.058x**	**0.107x**
	Base: w/ 4RFF	1.221x	1.628x	0.329x	0.384x	0.929x	0.524x	0.744x	0.283x	0.062x	0.115x
	w/ 6RFF	1.255x	1.660x	**0.302x**	0.376x	**0.857x**	0.524x	0.692x	0.273x	0.060x	0.108x
	w/ 6RFF, Noise	1.261x	1.679x	0.312x	0.373x	**0.857x**	**0.476x**	0.692x	0.263x	**0.058x**	0.108x

threshold of Z meters. The percentile metrics (p25, p50, p95) capture the distribution of error distances (actual vs predicted geocode in meters) on the test set. Overall, all AddressBind based models improve over baselines by a good margin for both geographies except *AddressBind-4RFF* variant falls short by a little in comparison to *RoBERTa-Triplet-H3* in accuracy metrics for USTR, which can be attributed to lower capacity of its location encoder module and unique challenges posed by addresses in unstructured geographies as highlighted in Section 1. However, *AddressBind-6RFF-4xNoise* outperforms SOTA sentence embedding baseline *Multilingual-E5-large-Instruct* as well as address embedding baseline *RoBERTa-Triplet-H3*. Improvement in precision by 7.7% and reduction in DR by 16% in the STR (4.3%, 13% for USTR) are observed over *RoBERTa-Triplet-H3*. Further, superior performance of *AddressBind-6RFF-4xNoise* over *AddressBind-4RFF* and *AddressBind-6RFF* validates the importance of having multiple RFFs with varying frequencies in geocode encoder and utilizing extra negatives while training.

4.3 Anomalous Geocode Detection

In this section, we demonstrate the effectiveness of our cross-modal approach on the task of anomalous geocode detection. This is a generic task, which deals with detecting whether a geocode associated with an address is valid or not. It manifests in multiple applications such as confirming if a delivery happened at the desired location and predicting the correctness of a learnt geolocation. As we embed geocodes and addresses in the same high dimensional space, we show that a simple cosine similarity operation with a carefully chosen threshold Θ can be highly effective in identifying anomalous geocodes. We create a test set of 20K addresses unseen during AddressBind model training and sample 4 geocodes at varying error distances (spanning up to several kms) from the actual distance of an address. We perform stratified sampling to maintain a fair balance of geocodes nearby and faraway to addresses' actual locations (cf. Figure 6). A validation set is used to determine the best individual cosine similarity threshold Θ for anomalous geocodes detection at different error distances. Table 4 reports the macro-averaged performance metrics at various error distances where x denotes a small distance deviation from actual location and further different multipliers to x conveys larger error deviations[5]. It can be seen that the model achieves up to F1 score of 0.94 (STR) and 0.92 (USTR) in classifying if a geocode is within certain error distance of an address or not. Refer to Section 5.2 for a qualitative analysis and precision-recall curves at various error distances. We observe that our model performance peaks at 10x and then go slightly down for higher errors, which can be potentially due to the increasing search space at higher error distances. Overall, the encouraging results demonstrate effectiveness of AddressBind to cross-modal tasks.

Table 4: Performance metrics on detecting anomalous geocodes at various error distances

Error Dist	Precision		Recall		F1		AUC	
	STR	USTR	STR	USTR	STR	USTR	STR	USTR
x	0.52	0.54	0.63	0.59	0.50	0.55	0.77	0.84
5x	0.84	0.81	0.86	0.85	0.83	0.82	0.90	0.91
10x	**0.94**	**0.92**	**0.94**	**0.92**	**0.94**	**0.92**	**0.98**	**0.96**
100x	0.92	0.84	0.91	0.80	0.91	0.82	**0.98**	0.94
200x	0.90	0.84	0.90	0.78	0.90	0.81	0.97	0.93

4.4 Latency and Scaling Efficiency Analysis

We conduct extensive latency and throughput measurements across different models to evaluate their practical applicability in real-world production scenarios. Table 5 presents the inference performance across various batch sizes (1, 64, and 1024) and Figure 4 visualizes performance characteristics in log space at multiple batch sizes varying from 1 to 1024, revealing substantial differences between model architectures to inform deployment decisions. All benchmarks were conducted using a single consumer-grade GPU to provide realistic performance metrics for typical deployment scenarios.

The proposed AddressBind model, with only 22.7M parameters and embedding dimension of 384, achieves the fastest inference times at 5.1 ± 0.1 ms for single queries. This performance makes it particularly suitable for latency-sensitive applications requiring real-time responses. The model maintains high throughput scaling to 8713.2 ± 481.5 samples/second at batch size 1024, demonstrating efficient utilization of GPU compute resources even under high load conditions. The RoBERTa-based models, with moderate parameter counts (83.1M) and embedding dimension 768, demonstrate comparable single-query latency (5.2 ± 0.1 ms) but exhibit less efficient scaling at higher batch sizes. This plateauing effect becomes particularly evident in batch scenarios beyond 64 samples.

[5] Metrics thresholds are masked and error distances are reported relative due to business confidentiality reason.

The Multilingual-E5-Large-Instruct model represents the relatively heavyweight option in our evaluation, with 559.9M parameters and embedding dimension 1024. It demonstrates substantially higher latency even for single queries (16.7 ± 0.1 ms), which increases to 1977.4 ± 76.9 ms at batch size 1024, nearly 17 times slower than AddressBind at equivalent batch size. Its throughput capacity is similarly constrained, reaching a maximum of only 558 samples/second before declining at the highest batch sizes, suggesting memory bandwidth or other computational bottlenecks.

The top panel in Figure 4 illustrates approximately linear latency growth in log-log space across all models with AddressBind being the lowest curve. The throughput visualization in the bottom panel demonstrates that all models initially benefit from increased batch sizes, but with dramatically different saturation points. While AddressBind models continue scaling effectively to the largest tested batch size, RoBERTa models plateau beyond batch size 64, and the Multilingual-E5-Large-Instruct model saturates at under 600 samples/second with slight performance regression at the highest batch sizes. This combination of low latency, high throughput and better embedding quality (cf. Section 4.1, 4.2) positions AddressBind as an ideal candidate for production deployment across diverse operating conditions.

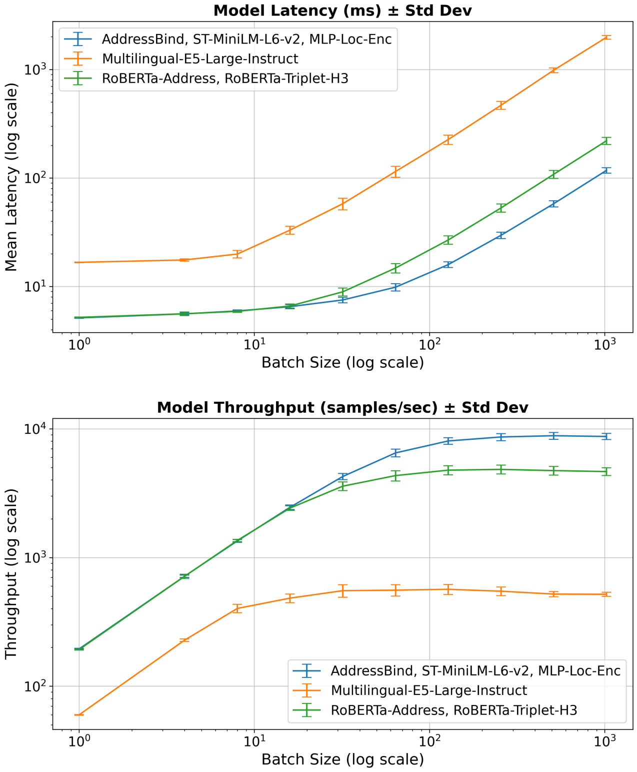
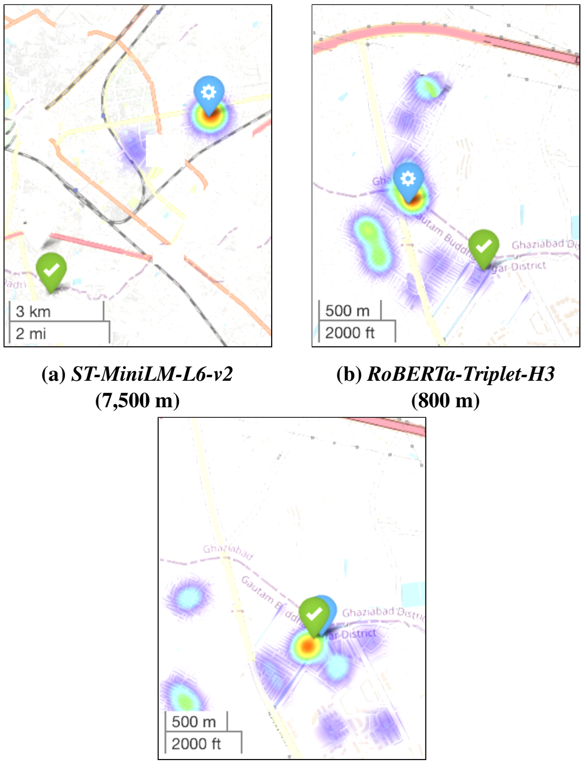


Figure 4: Latency (milliseconds) and throughput curves (samples per second) with error bars in log space for different models at various batch sizes

5 Qualitative Analysis and Ablations

5.1 Anecdotes on Address Geocoding

Figure 5 illustrates the geocoding process for address *X - block, XX, Yatender nagar, Tigri Gautam Budh Nagar, up, Gazibad, 201009, Ghaziabad, UP, IN* using different models. This address poses an interesting challenge because it lies on the jurisdiction boundary of two cities, and ambiguity in usage of the two city names may arise. Here, both *Ghaziabad* and *Gautam Budh Nagar* are mentioned in address lines and latter is the correct choice as can be seen in map visuals below. We observe that *ST-MiniLM-L6-v2* model gets completely perplexed and produces very far away addresses as retrieved neighbours. *RoBERTa-Triplet-H3* improves over the previous model but is still not able to disambiguate the address text. In contrast, we can see that our proposed *AddressBind-6RFF-4xNoise* model demonstrates superior understanding by fetching multiple addresses from the true neighbourhood of the query address despite the subtlety.



(a) ***ST-MiniLM-L6-v2*** **(7,500 m)**

(b) ***RoBERTa-Triplet-H3*** **(800 m)**

(c) ***AddressBind-6RFF-4xNoise*** **(50 m)**

Figure 5: Retrieved nearest neighbours (depicted via heap map) with predicted and actual locations, and error distances from actual location by various models (Note: Background maps are intentionally morphed to preserve privacy)

Table 5: Models architecture specifications and scaling efficiency performance comparison

Model Group	Params	Dims	Latency (ms) ± std			Throughput (samples/s) ± std		
			Batch=1	Batch=64	Batch=1024	Batch=1	Batch=64	Batch=1024
AddressBind (ours) ST-MiniLM-L6-v2 MLP-Loc-Enc	22.71M	384	**5.1** ± 0.1	**9.9** ± 0.7	**117.5** ± 6.9	**194.6** ± 2.1	**6494.8** ± 457.1	**8713.2** ± 481.5
RoBERTa-Address RoBERTa-Triplet-H3	83.12M	768	5.2 ± 0.1	14.8 ± 1.5	220.2 ± 16.7	192.4 ± 1.9	4328.4 ± 393.4	4650.3 ± 328.0
Multilingual-E5-Large-Instruct	559.9M	1024	16.7 ± 0.1	114.8 ± 13.1	1977.4 ± 76.9	59.9 ± 0.4	557.6 ± 57.3	517.8 ± 19.4

5.2 Anecdotes on Anomalous Geocode Detection

Figure 6 illustrates the spatial distribution of cosine similarity between an address and surrounding geocodes. The visualization reveals a clear gradient pattern where similarity decreases with increasing distance from the actual location. This demonstrates the model's ability to effectively encode spatial proximity in the shared embedding space. Notably, the similarity values exhibit a non-linear decay, with steeper drops occurring within the first few hundred meters, suggesting the model has learned to be particularly sensitive to small displacements in dense urban environments. Further, a distinctive "+" shaped light colored pattern emerges around the actual location (indicated by dashed lines for visual guidance), which corresponds to the intersection of crossing roads (faintly shown in the background). This pattern is significant as it hints how the model may implicitly learn to some extent the underlying road network structure without explicit supervision. The higher similarity values along these road axes reflect the model's understanding that addresses are more likely to be situated along transportation corridors rather than in arbitrary locations.

Figures 7 presents precision-recall curves for anomalous geocode detection across various error thresholds for structured geographies. The performance progression follows a non-monotonic pattern, peaking at 10x with an AUC of 0.98 before slightly declining at larger distances. The 10x curve demonstrates near-perfect classification capability with both precision and recall above 0.94, suggesting an optimal detection threshold for practical applications. Curves for smaller error distances exhibit significantly lower performance with more pronounced trade-offs, indicating the challenge of distinguishing minor geocode deviations. The convergence at higher error distances suggests diminishing returns beyond certain spatial thresholds, possibly due to increasing sparsity of relevant contextual information at larger scales.

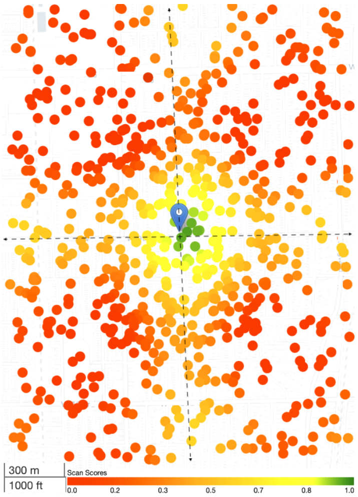


Figure 6: Cosine similarity between embeddings of nearby geocodes and address *XXXX S Bell Ave, 60643, Chicago, IL, US*

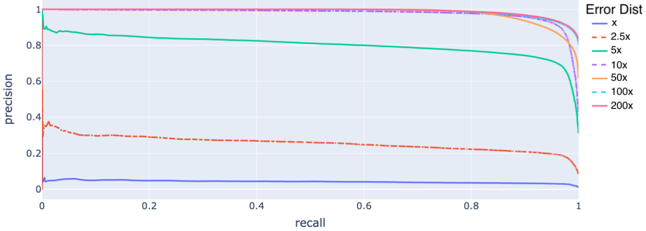


Figure 7: PR curves for anomalous geocode detection at different error distances

5.3 Ablations on Random Fourier Features

Learning Curves with/without Random Fourier Features Figure 8 depicts train loss curve to demonstrates the critical role of Random Fourier Features in model convergence. Without RFF projections, the MLP-Loc-Enc model exhibits pronounced oscillations in both training and validation losses, failing to converge even after 50 epochs. This instability stems from the inherent difficulty neural networks face when mapping low-dimensional geocode inputs directly to high-dimensional semantic spaces. In contrast, models employing RFF projections display stable learning curves with faster convergence (i.e., curve on the bottom of Figure 8). This empirical evidence strongly supports the theoretical understanding that RFF projections provide essential inductive bias for learning high-frequency functions from low-dimensional spatial inputs.

Impact of σ on RFF Sensitivity Figure 9 illustrates the impact of σ value on RFF projections for a pair of geocodes, which are 200 meters apart. The two rows in each of the sub-figures represent the projected vectors for geocodes at certain value of hyperparameter σ, and intensity of the color depicts the magnitude of cell value. It can be observed that larger values of σ result in higher sensitivity of an

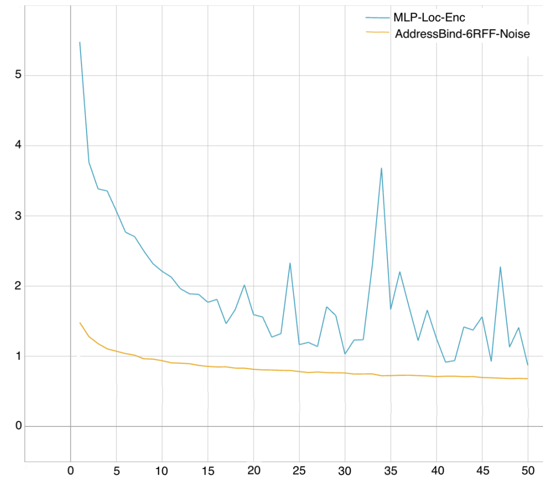


Figure 8: Train loss curves of models with and without RFF projections over the course of 50 epochs

RFF projection module towards fine-grained changes in geocodes (i.e., lesser correlation in cell colors across the two rows).

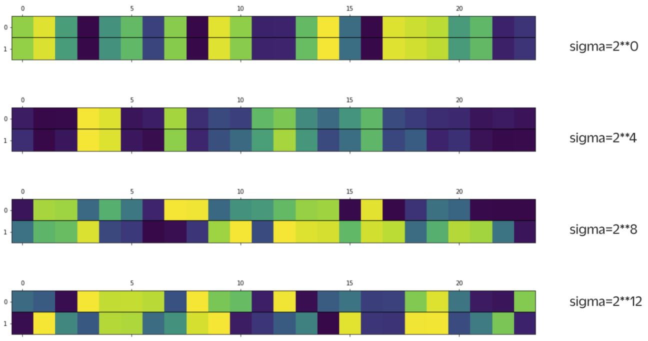


Figure 9: Illustration on impact of σ value on RFF projections for a pair of geocodes, which are 200m away

6 Conclusion

In this paper, we have introduced a novel multi-modal approach to learn proximity-aware quality address representations by embedding both geocodes and customer addresses in the same space, and eliminate dependency on triplets engineering. Our extensive experimentation shows significant gains in learnt embeddings quality intrinsically and on address geocode learning over the SOTA models for instruction-tuned sentence embeddings as well as address focused embeddings. Further, the proposed model shows potential for cross-modal applications such as detecting anomalous geocodes w.r.t. addresses. Significant improvements shown in geocoding accuracy and reduced defects for addresses directly translate to improved performance for systems in logistics and related domains.

References

[1] 2025. MTEB: Leaderboard for Massive Multilingual Text Embedding Benchmark. https://huggingface.co/spaces/mteb/leaderboard. [Online; accessed 25-May-2025].

[2] Tom Patterson Bojan Šavrič and Bernhard Jenny. 2019. The Equal Earth map projection. *International Journal of Geographical Information Science* 33, 3 (2019), 454–465. doi:10.1080/13658816.2018.1504949 arXiv:https://doi.org/10.1080/13658816.2018.1504949

[3] Ben Carterette and Ellen M. Voorhees. 2011. *Overview of Information Retrieval Evaluation*. Springer Berlin Heidelberg, Berlin, Heidelberg, 69–85. doi:10.1007/978-3-642-19231-9_3

[4] Brandon Clark, Alec Kerrigan, Parth Parag Kulkarni, Vicente Vivanco Cepeda, and Mubarak Shah. 2023. Where We Are and What We're Looking At: Query Based Worldwide Image Geo-localization Using Hierarchies and Scenes. arXiv:2303.04249 [cs.CV]

[5] Kenneth Enevoldsen et al. 2025. MMTEB: Massive Multilingual Text Embedding Benchmark. arXiv:2502.13595 [cs.CL] https://arxiv.org/abs/2502.13595

[6] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821* (2021).

[7] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. ImageBind: One Embedding Space To Bind Them All. arXiv:2305.05665 [cs.CV]

[8] Govind, Céline Alec, Jean-Luc Manguin, and Marc Spaniol. 2021. FETD^2: A Framework for Enabling Textual Data Denoising via Robust Contextual Embeddings. In *Linking Theory and Practice of Digital Libraries*, Gerd Berget, Mark Michael Hall, Daniel Brenn, and Sanna Kumpulainen (Eds.). Springer International Publishing, Cham, 3–16.

[9] Govind and Saurabh Sohoney. 2022. Learning Geolocations for Cold-Start and Hard-to-Resolve Addresses via Deep Metric Learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Yunyao Li and Angeliki Lazaridou (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 322–331. doi:10.18653/v1/2022.emnlp-industry.33

[10] James Hays and Alexei A. Efros. 2008. IM2GPS: estimating geographic information from a single image. *2008 IEEE Conference on Computer Vision and Pattern Recognition* (2008), 1–8. https://api.semanticscholar.org/CorpusID:2061602

[11] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

[12] Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. 2023. SatCLIP: Global, General-Purpose Location Embeddings with Satellite Imagery. *arXiv preprint arXiv:2311.17179* (2023).

[13] Kartik Kuckreja, Muhammad S. Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad S. Khan. 2024. GeoChat: Grounded Large Vision-Language Model for Remote Sensing. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024).

[14] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) *(ICML'23)*. JMLR.org, Article 814, 13 pages.

[15] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. arXiv:2201.12086 [cs.CV]

[16] Sha Li, Chao Zhang, Dongming Lei, Ji Li, and Jiawei Han. 2019. *GeoAttn: Localization of Social Media Messages via Attentional Memory Network.* Proceedings of the 2019 SIAM International Conference on Data Mining (SDM), 64–72. doi:10.1137/1.9781611975673.8 arXiv:https://epubs.siam.org/doi/pdf/10.1137/1.9781611975673.8

[17] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. doi:10.48550/ARXIV.1907.11692

[18] Saket Maheshwary and Saurabh Sohoney. 2023. Learning Geolocation by Accurately Matching Customer Addresses via Graph based Active Learning. In *Companion Proceedings of the ACM Web Conference 2023* (Austin,TX,USA) *(WWW '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 457–463. doi:10.1145/3543873.3584647

[19] Gengchen Mai, Ni Lao, Yutong He, Jiaming Song, and Stefano Ermon. 2023. CSP: self-supervised contrastive spatial pre-training for geospatial-visual representations. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu,Hawaii,USA) *(ICML'23)*. JMLR.org, Article 981, 18 pages.

[20] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2021. NeRF: representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (dec 2021), 99–106. doi:10.1145/3503250

[21] Vamsi Krishna Penumadu, Nitesh Methani, and Saurabh Sohoney. 2022. Learning geospatially aware place embeddings via weak-supervision. In *Proceedings of the*

30th International Conference on Advances in Geographic Information Systems (Seattle,Washington) *(SIGSPATIAL '22)*. Association for Computing Machinery, New York, NY, USA, Article 80, 10 pages. doi:10.1145/3557915.3561016

[22] Chunyao Qian, Chao Yi, Chengqi Cheng, Guoliang Pu, and Jiashu Liu. 2020. A coarse-to-fine model for geolocating chinese addresses. *ISPRS International Journal of Geo-Information* 9, 12 (2020), 698.

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV]

[24] Ali Rahimi and Benjamin Recht. 2007. Random features for large-scale kernel machines. In *Proceedings of the 20th International Conference on Neural Information Processing Systems* (Vancouver, British Columbia, Canada) *(NIPS'07)*. Curran Associates Inc., Red Hook, NY, USA, 1177–1184.

[25] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[26] Marc Rußwurm, Konstantin Klemmer, Esther Rolf, Robin Zbinden, and Devis Tuia. 2024. Geographic Location Encoding with Spherical Harmonics and Sinusoidal Representation Networks. ICLR. arXiv:2310.06743

[27] Bhavuk Singhal, Anshu Aditya, Lokesh Todwal, Shubham Jain, and Debashis Mukherjee. 2024. GeoIndia: A Seq2Seq Geocoding Approach for Indian Addresses. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Franck Dernoncourt, Daniel Preoţiuc-Pietro, and Anastasia Shimorina (Eds.). Association for Computational Linguistics, Miami, Florida, US, 395–407. doi:10.18653/v1/2024.emnlp-industry.29

[28] Vishal Srivastava, Priyam Tejaswin, Lucky Dhakad, Mohit Kumar, and Amar Dani. 2020. A Geocoding Framework Powered by Delivery Data. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems* (Seattle, WA, USA) *(SIGSPATIAL '20)*. Association for Computing Machinery, New York, NY, USA, 568–577. doi:10.1145/3397536.3422254

[29] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. 2020. Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. arXiv:2006.10739 [cs.CV]

[30] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748 [cs.LG] https://arxiv.org/abs/1807.03748

[31] Vicente Vivanco, Gaurav Kumar Nayak, and Mubarak Shah. 2023. GeoCLIP: Clip-Inspired Alignment between Locations and Images for Effective Worldwide Geo-localization. In *Advances in Neural Information Processing Systems*.

[32] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 Text Embeddings: A Technical Report. arXiv:2402.05672 [cs.CL] https://arxiv.org/abs/2402.05672

[33] Ellen D. Zhong, Tristan Bepler, Joseph H. Davis, and Bonnie Berger. 2020. Reconstructing continuous distributions of 3D protein structure from cryo-EM images. arXiv:1909.05215 [q-bio.QM]

[34] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Wang HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Cai Wan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. 2023. LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment. arXiv:2310.01852 [cs.CV]