Semantic Ambiguity Modeling and Propagation for Fine-Grained Visual Cross View Geo-Localization

Mingtao Feng[1], Fenghao Tian[1], Jianqiao Luo[2*], Zijie Wu[2]
Weisheng Dong[1], Yaonan Wang[2], Ajmal Mian[3]
[1]Xidian University
[2]Hunan University
[3]University of Western Australia

Abstract

Visual cross view geo-localization is generally approached within a joint retrieval-and-calibration framework. However, existing methods overlook semantic ambiguities arising from query and reference images characterized by low overlap, dynamic foregrounds, viewpoint changes, and perceptual aliasing. This makes it challenging to automatically control the relative importance of the two tasks, potentially compromising the retrieval task in favor of the offset regression. Consequently, the model may encounter conflicting dominating gradients during joint training. To address this, we propose to model the semantic ambiguity during the offset regression process by integrating associated uncertainty scores, represented as 2D Gaussian distributions, to mitigate negative transfer effects within the joint tasks. We further introduce an uncertainty-aware similarity metric to enhance similarity assessment between query and reference images, accounting for their semantic ambiguities. This metric propagates uncertainty scores into the retrieval task, focusing on certain samples and learning discriminative feature embeddings, allowing the model to adaptively handle conflicting dominating gradients during joint training. Extensive experiments demonstrate that our method improves the overall performance of the joint tasks, achieving state-of-the-art results on the VIGOR and CVACT datasets.

Code — https://github.com/Afoolbird/SAMP

Introduction

The goal of fine-grained visual cross view geo-localization is to predict the geographic location of a ground query image with respect to a GPS-tagged reference image database. Utilizing readily available reference images as maps offers a cost-effective and promising approach to localization, with significant potential for applications such as autonomous driving (Wang et al. 2024), ant robot navigation (Hou et al. 2024; Feng et al. 2023b,a, 2022, 2021), etc.

Previous works (Yang et al. 2021; Liu and Li 2019; Zhu et al. 2022; Xia et al. 2023; Li et al. 2024) assume ***ideal scenarios***: 1) each ground query image has a corresponding aerial reference image precisely centered on the query location, using the GPS coordinate of the central pixel for localization; 2) the matching reference image is already available with coarse GPS priors, requiring only offset regression without global image retrieval. However, these assumptions are far from ***practical scenarios*** because: 1) query images can be captured at arbitrary locations in the area of interest, with reference images typically taken before the queries arise, leading to misalignments and location estimation errors of tens of meters when query images are far from the reference image center. 2) In cases of GPS denial or large GPS errors, coarse GPS priors cannot reliably select matching reference images, making fine-grained regression alone insufficient for accurate localization. VIGOR (Zhu et al. 2021a) first identifies the issue of reference samples that, while not centered on query locations, partially cover them, leading to multiple overlapping reference images (yellow boxes in Fig.1a) for the same query location (red stars in Fig.1a). This disrupts the one-to-one correspondence, necessitating image retrieval for *coarse* localization and within-image calibration for *fine-grained* offset prediction. Consequently, the cross view geo-localization problem is generally addressed within a joint retrieval-and-calibration framework, where retrieval and offset regression objectives are

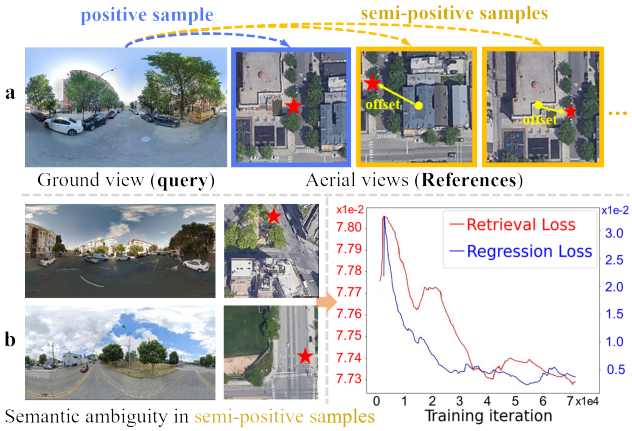

Figure 1: a) Positive sample (query image and blue box aerial image) and semi-positive samples (query image and yellow box aerial image). b) Semantic ambiguity destabilizes the training process, resulting in negative transfer within the joint task.

*Corresponding author.

optimized jointly from a shared feature representation in an end-to-end manner.

While the joint retrieval-and-calibration framework (Zhu et al. 2021a) leverages supervision from semi-positive samples, simply treating these samples as positives introduces semantic ambiguity, undermining both retrieval learning and offset regression performance. This is because semi-positive aerial images capture only a small portion of the query scene, and their feature embedding similarities with the query should not be as high as those of fully positive samples. In Fig.1b, dynamic foreground elements (e.g., pedestrians and vehicles), viewpoint changes, and perceptual aliasing (Lowry et al. 2015) (hard to differentiate query location from similar aerial images) in semi-positive samples exacerbate appearance gaps and increase semantic ambiguity during offset regression. Directly optimizing an objective averaged across joint tasks without accounting for these issues can result in imbalanced task prioritization, potentially compromising retrieval performance in favor of offset regression, e.g., forcing the model to regress the offset for semi-positive samples will cause a negative transfer problem, degrading the retrieval performance, as depicted in Fig.1b. Consequently, the model may encounter conflicting dominating gradients (Yu et al. 2020), destabilizing the training process and overall performance. However, existing methods do not address the semantic ambiguity challenges in visual cross view geo-localization for practical scenarios.

In this paper, we propose a principled approach to optimally balance the retrieval and offset regression tasks, achieving superior performance compared to a naive weighted sum of the two tasks. We first model the semantic ambiguity of semi-positive samples by estimating the Cholesky coefficients of the covariance matrix through a 2D Gaussian distribution for global image features of query and reference samples. These estimates are integrated using a Gaussian log-likelihood loss function, allowing simultaneous estimation of the query offset and its associated uncertainty score. This score is then used to reduce the impact of offset regression loss in cases where high uncertainty is observed in predicted offsets of semi-positive samples, effectively mitigating the negative transfer effects in joint tasks caused by semi-positive samples. Furthermore, we introduce an uncertainty-aware similarity metric to facilitate similarity assessments between query and reference images, accounting for their semantic ambiguities. This metric propagates uncertainty scores into the retrieval task, enabling the model to focus on certain samples and learn discriminative feature embeddings. This approach allows the joint tasks to address semantic ambiguity effectively, facilitating adaptive learning and mitigating the impact of conflicting gradients during training. Our main contributions are:

- We address the semantic ambiguity issues arising from semi-positive samples in visual cross view geo-localization, and propose a framework that automatically balances the importance of retrieval and calibration.
- We model the semantic ambiguity during the offset regression process by integrating associated uncertainty scores, aimed at mitigating negative transfer effects within the joint tasks of retrieval and calibration.
- We introduce an uncertainty-aware similarity metric that propagates uncertainty scores into the retrieval task, enabling the model to adaptively manage conflicting dominating gradients during joint training.

Extensive experiments on the VIGOR and CVACT datasets show that our method improves the overall performance of the joint tasks, achieving state-of-the-art results.

Related Works

Visual cross view geo-localization. The significant appearance gap between the two views and poor metric learning techniques make it challenging for visual cross view localization (Dai et al. 2021; Hu et al. 2023). Efficient network architectures, e.g. polar transformation (Shi et al. 2022) and generative networks (Toker et al. 2021), are introduced to reduce the domain gap. The spatial-aware feature aggregation (SAFA) (Shi et al. 2019) or vision transformer (Zhu et al. 2022; Dai et al. 2023; Lu et al. 2024) modules are also proposed to enhance the discriminability of features for the retrieval task. In addition, differentiable Lukas-Kanade optimizers are proposed to iteratively compute the relative pose between the query and reference images (Cao et al. 2023; Zhang et al. 2023b). Unlike them, we are the first to consider semantic ambiguity issues of the appearance gap for ***practical scenarios*** in visual cross view geo-localization, proposing a novel framework to improve the overall performance.

Multi task learning. Different tasks within a unified framework occupy diverse learning spaces, and the lack of control in joint training can result in a negative transfer problem. Conflicting gradients with opposing directions (Chai et al. 2022) and dominating gradients by a certain task (Senushkin et al. 2023) during training may hinder the search for a Pareto stationary solution (Senushkin et al. 2023). Current uncertainty-weighted task balancing methods treat task-level uncertainty as fixed network parameters, uniformly weighting all samples within a task, disregarding data variations (Chen et al. 2022). In the joint retrieval-and-calibration geo-localization framework, modeling uncertainty and automatically balancing task importance in an end-to-end manner is challenging. To address this, we treat query offset uncertainty as dynamic network outputs rather than fixed parameters, accommodating semantic ambiguity across samples and providing confidence in localization quality during inference. These uncertainty scores are then propagated into the retrieval task to adaptively manage conflicting and dominant gradients during joint training, enhancing overall performance.

Method

Each dataset unit includes the images I and annotation y^*:

$$\{I_g,\ I_p,\ I_{s_{\{1,\cdots,m\}}},\ y_p^*,\ y_{s_{\{1,\cdots,m\}}}^*\}, \tag{1}$$

where I_g is the ground view query image, the aerial view images I_a include the corresponding positive reference image I_p and semi-positive reference images I_{s_i}. The semi-positive reference image number m is set to 3 (Zhu et al.

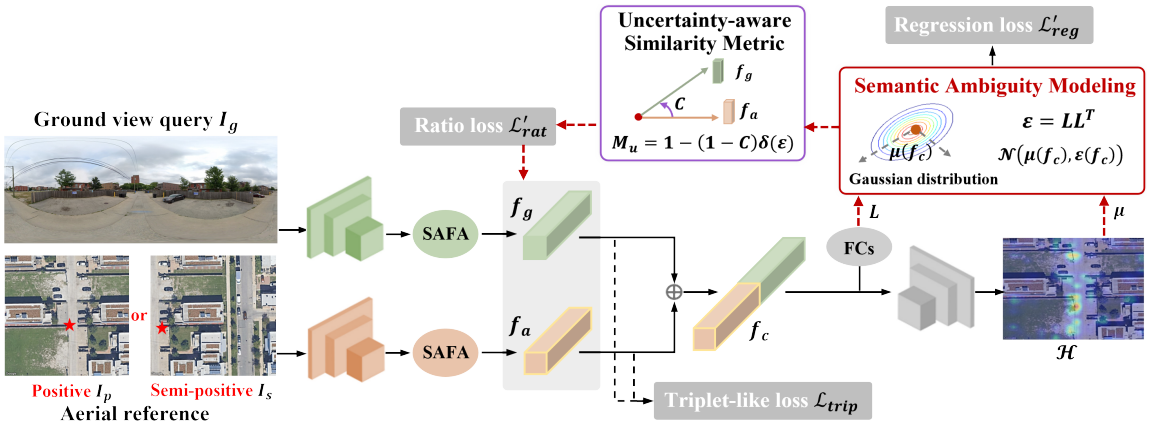

Figure 2: The overview of our proposed framework. It consists of three parts: (1) learning of retrieval task through the triplet-like and ratio losses, (2) semantic ambiguity modeling in the offset regression task, and (3) propagation of the ambiguity to the ratio loss to improve joint training.

2021a). $y^* \in R^2$ denotes the latitude and longitude offset of the query location with respect to the reference image center.

The joint retrieval-and-calibration framework employs both positive and semi-positive samples: 1) the retrieval task operates within the discriminative feature embedding space to match query and reference images. 2) the regression task predicts the offset for the query location within the aerial reference image. The total loss is:

$$\mathcal{L} = \underbrace{\mathcal{L}_{trip} + \mathcal{L}_{rat}}_{Task1} + \underbrace{\mathcal{L}_{reg}}_{Task2}, \tag{2}$$

where $\mathcal{L}_{trip}$ denotes the triplet-like loss and $\mathcal{L}_{rat}$ denotes the ratio loss. The first retrieval task is supervised by $\mathcal{L}_{trip}$ and $\mathcal{L}_{rat}$ concurrently. $\mathcal{L}_{reg}$ represents the second offset regression task loss.

The objective of the first retrieval task is to acquire a feature embedding space where image features of positive and semi-positive samples can be effectively aligned. Considering the positive samples, we choose the InfoNCE loss (Deuser et al. 2023) as $\mathcal{L}_{trip}$ to guarantee the similarity of feature embeddings among them:

$$\mathcal{L}_{trip} = -\log \frac{\exp(C_p/\tau))}{\sum_{i=0}^{N} \exp(C_i/\tau))}, \tag{3}$$

where C_p denotes the feature similarity within positive samples, τ is a temperature parameter, and N is the batch size. For each batch, we adopt a hard sample mining strategy (Deuser et al. 2023) to make full use of all negative samples.

For semi-positive samples, the images I_g and I_s characterized by low overlap rates can also benefit the learning of the feature embedding (Zhu et al. 2021a). An intuitive idea is to assign the embedding similarity according to the Intersection of Union (IOU) between the query and reference images, the ratio loss $\mathcal{L}_{rat}$ can be formulated as:

$$\mathcal{L}_{rat} = \left[\frac{O_s}{O_p} - \frac{C_s}{C_p}\right]_+, \tag{4}$$

where O_p and O_s denote the IOU of positive and semi-positive samples, C_p and C_s denote the features cosine similarity of the positive and semi-positive samples, respectively.

For the second task, the fine-grained geo-localization is achieved by regressing the query offset $\tilde{y}$ in the reference image I_s. Given the predicted offset $\tilde{y}$, the offset regression task loss $\mathcal{L}_{reg} = \| \tilde{y} - y^* \|_2$. It overlooks the semantic ambiguity introduced by semi-positive samples, hindering the effective learning of query offsets during joint training. This limitation results in negative transfer issues and overall performance degradation. Additionally, ignoring semantic ambiguity in the retrieval task impairs feature embedding learning, as ambiguous samples introduce confusion into the network (Wei et al. 2021). Motivated by this, we model semantic ambiguity in $\mathcal{L}_{reg}$ to mitigate negative transfer between retrieval and regression tasks and introduce a novel similarity metric in feature embedding space to propagate this ambiguity to the retrieval task, as shown in Fig.2.

Semantic Ambiguity Modeling

Given the query image I_g and aerial reference images I_a, they go through a siamese backbone and SAFA blocks (Shi et al. 2019) to generate the global image features f_g and f_a for the ground view image and aerial images respectively, which are used for $\mathcal{L}_{trip}$ and $\mathcal{L}_{rat}$ to learn the retrieval task. We then concatenate the embedded features f_g and f_a as the image fusion features f_c, which is fed into a decoder to generate a heatmap $\mathcal{H}$, $\mathcal{H} \in R^{H_a \times W_a}$. Element $\mathcal{H}(y)$ represents the probability that the query camera is located at position y in a $H_a \times W_a$ grid. $\mathcal{H}$ is normalized by 2D Softmax function (Kumar et al. 2019), ensuring $\sum_y \mathcal{H}(y) = 1$. The offset $\tilde{y}$ is estimated as the spatial mean over $\mathcal{H}$, $\tilde{y} = \sum y\mathcal{H}(y)$. Notably, the spatial mean operation over $\mathcal{H}$ is differentiable, enabling end-to-end optimization. This approach is more effective than the multi-perceptual layer prediction strategy in VIGOR for offset regression, as neural networks often struggle with continuous value prediction (Xia et al. 2022).

We then represent the uncertainty of the query offset as a Gaussian distribution with covariance matrix ε, which is a

2×2 symmetric positive definite matrix with three degrees of freedom. Accordingly, we use the fusion feature f_c to learn the lower triangular matrix L for ε using Cholesky decomposition, i.e. $LL^{\mathbf{T}} = \varepsilon$. f_c passes through four fully connected layers to derive the three variables determining L, and the exponential linear unit (ELU) activation function (Kumar et al. 2019) is used to ensure L has positive diagonal elements. We learn ε from image fusion features because f_c captures the global semantics of cross view images, addressing semantic ambiguity and uncertainty through comparative analysis of two images. Recent fine-grained geo-localization methods use possibility heatmaps (Fervers et al. 2023) to manage offset uncertainty, which can lead to overconfidence in localization. In contrast, we independently learn ε from f_c in parallel with the decoder, providing a more reliable method for modeling semantic ambiguity.

The regression task predicts a data-dependent 2D Gaussian distribution $\mathcal{N}(\mu(f_c), \varepsilon(f_c))$ over the offset, where $\mu(f_c)$ is set as the predicted offset $\tilde{y}$. The regression task is learned through optimization of the likelihood associated with the ground truth offset $P(y^*|\mu(f_c), \varepsilon(f_c))$,

$$P(y^*|\mu(f_c), \varepsilon(f_c)) = \mathcal{N}(\mu(f_c), \varepsilon(f_c)). \tag{5}$$

Thus, the new regression loss $\mathcal{L}'_{reg}$ is defined as:

$$\begin{aligned}\mathcal{L}'_{reg} &= -\ln P(y^*|\mu(f_c), \varepsilon(f_c)) \\ &= (y^* - \mu(f_c))^{\mathbf{T}} \varepsilon(f_c)^{-1} (y^* - \mu(f_c)) + \ln|\varepsilon(f_c)| + A,\end{aligned} \tag{6}$$

where $||$ denotes determinant and A is a constant equal to $-0.5\ln(2\pi)$. The first component penalizes the discrepancy between the predicted mean and the true offset, while the second component regularizes high uncertainty.

Guided by Eq. 6, the joint framework mitigates distraction from ambiguous images, reducing negative transfer between retrieval and regression tasks. In cases of ambiguous samples with challenging offset prediction, the network allows a larger magnitude for ε, signaling high uncertainty to moderate the severe residual $(y^* - \mu(f_c))$. The impact of ambiguous samples during training is thereby minimized. This mechanism functions as a loss attenuation strategy, reducing the influence of the regression task in the presence of ambiguous inputs and preventing degradation in retrieval performance.

Uncertainty-aware Similarity Metric

Given the query and reference feature embeddings f_g and f_a, we tune the degree of their cosine similarities $C_{(s/p)}(f_g, f_a)$ according to the semantic ambiguity, to prevent the network from over-optimizing the similarity for ambiguous images. Conditioned on the cosine similarity C, we define an uncertainty-aware similarity metric M_u to be tuned by uncertainty ε, which is formulated as:

$$\begin{aligned} M_u &= 1 - (1 - C)\delta(\varepsilon), \\ \delta(\varepsilon) &= \frac{1}{1 + \alpha e^{\beta(|\varepsilon(f_c)| - \gamma)}}, \end{aligned} \tag{7}$$

where $\delta(\varepsilon)$ is a decaying exponential function (Kumar et al. 2019), including three positive parameters α, β, and γ. The value of $\delta(\varepsilon)$ ranges from 0 to 1 (see supplementary material for its function curves). The increase in uncertainty reduces $\delta(\varepsilon)$, causing M_u to approach its upper bound of 1.0. This prevents the network from incorrectly emphasizing embedding similarity between ambiguous images. Conversely, low uncertainty results in a small M_u, prompting the network to enhance similarity for certain images.

The semantic ambiguity is mainly caused by the semi-positive samples merged in ratio loss $\mathcal{L}_{rat}$. Equipped with the proposed M_u, we design a new ratio loss $\mathcal{L}'_{rat}$:

$$\mathcal{L}'_{rat} = \left[\frac{O_s}{O_p} - \frac{M_{us}}{M_{up}}\right]_+, \tag{8}$$

where M_{us} and M_{up} denote the embedding similarity of semi-positive and positive samples, respectively. By emphasizing certain samples, $\mathcal{L}'_{rat}$ takes full advantage of semi-positive samples to learn more discriminative features. This helps in retrieving the semi-positive reference images covering the query location, leading to better retrieval performance, especially in terms of hit rate.

As positive samples typically exhibit certainty regarding the query offset and consequently demonstrate very high similarity, the network focuses on enhancing embedding similarity for certain semi-positive samples. A gradient analysis is provided to demonstrate the effect of M_u in learning discriminate embedding space. Since we only substitute the simialrity metric in $\mathcal{L}_{rat}$ with M_u, we have $\frac{\partial \mathcal{L}_{rat}}{\partial C} = \frac{\partial \mathcal{L}'_{rat}}{\partial M_u}$, which is dependent on the form of loss function. Denoting the network parameters and feature embeddings as θ and f respectively, the loss gradient on parameter $\frac{\partial \mathcal{L}'_{rat}}{\partial \theta}$ can be deduced as:

$$\begin{aligned}\frac{\partial \mathcal{L}'_{rat}}{\partial \theta} &= \frac{\partial \mathcal{L}'_{rat}}{\partial M_u}\frac{\partial M_u}{\partial C}\frac{\partial C}{\partial f}\frac{\partial f}{\partial \theta} = \frac{\partial \mathcal{L}_{rat}}{\partial C}\frac{\partial M_u}{\partial C}\frac{\partial C}{\partial f}\frac{\partial f}{\partial \theta} \\ &= \frac{\partial \mathcal{L}_{rat}}{\partial \theta}\frac{\partial M_u}{\partial C} = \frac{\partial \mathcal{L}_{rat}}{\partial \theta}\delta(\varepsilon),\end{aligned} \tag{9}$$

where we can observe that $\frac{\partial \mathcal{L}_{rat}}{\partial \theta}$ is tuned by the decaying function $\delta(\varepsilon)$, i.e., higher uncertainty results in smaller gradients compared to the original gradient $\frac{\partial \mathcal{L}_{rat}}{\partial \theta}$, thereby exerting less influence on the network. Furthermore, the designed M_u enables the retrieval task to focus on salient landmarks shared by specific samples, providing more informative feature embeddings for the regression task. Both tasks address semantic ambiguity, reducing conflicting dominant gradients across the joint tasks, and ensuring aligned gradients during training (Senushkin et al. 2023), thereby contributing to a better optimum for both tasks.

Training and Inference

At the training stage, the proposed losses $\mathcal{L}'_{rat}$ and $\mathcal{L}'_{reg}$ are adopted to handle the semantic ambiguity for the retrieval and regression tasks respectively, and the joint framework is optimized by the new loss $\mathcal{L}'$:

$$\mathcal{L}' = \mathcal{L}_{trip} + \mathcal{L}'_{rat} + \mathcal{L}'_{reg}. \tag{10}$$

During inference, we use the modeled semantic ambiguity to select the optimal localization result from top-K candidate retrieved reference images. Given a query image, the

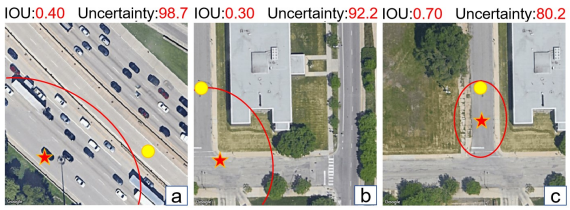

Figure 3: Visualization of the semantic ambiguity modeling. The yellow dots and red stars denote the GT and predicted localization. The uncertainty scores are shown in red circles.

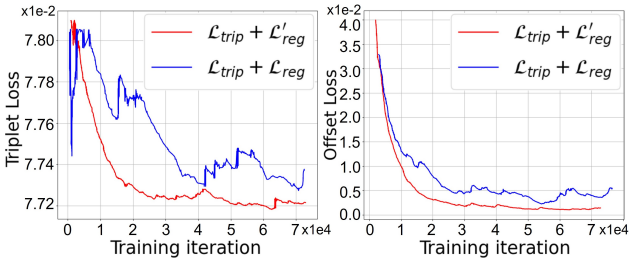

Figure 4: The comparisons of learning curves in terms of triplet-like loss and regression loss.

	$\mathcal{L}_{trip}$	$\mathcal{L}_{reg}$	$\mathcal{L}'_{reg}$	$\mathcal{L}_{rat}$	$\mathcal{L}'_{rat}$	k_0	Top-1	Top-5	Top-1%	Hit Rate	Mean-Error
1	✓						59.0	67.9	97.6	62.8	-
2	✓	✓					60.5	70.3	97.8	67.2	10.08
3	✓		✓				69.9	85.2	96.6	78.4	7.83
4	✓		✓	✓			73.0	87.7	97.4	81.8	6.25
5	✓		✓		✓		80.1	96.8	98.2	91.5	5.31
6	✓		✓		✓	✓	**80.6**	**96.8**	**98.2**	**93.9**	**4.91**

Table 1: Ablation studies on different losses. Experiments are conducted on the same area of the VIGOR dataset.

joint framework first provides top-K retrieved reference images $I_{ak}, k = 1, \cdots, K$, and then uses the feature embeddings to output the probability distribution of the query offset $\mathcal{N}(\mu_k, \varepsilon_k)$, where μ_k indicates the mean of the offset, and ε_k can be treated as the confidence evidence for the kth result, i.e., a small value of $|\varepsilon_k|$ suggests high-quality localization. Hence, the k_0th candidate with the lowest ambiguity is chosen as the optimal localization result, which is given by $k_0 = \arg\min |\varepsilon_k|$. By selecting the optimal k_0th result, i.e., the retrieved image I_{a,k_0} and its offset mean μ_{k_0} among the top-K retrieved results, we can achieve better retrieval performance over the original top-1 result.

Experiments

We use two benchmark datasets. **VIGOR:** Any arbitrary query location in the area of interest is covered by four reference images, a positive reference, and three semi-positive references. We follow the same-area and cross-area splits from (Zhu et al. 2021a). **CVACT:** Following the sampling strategy for ***practical scenarios*** in VIGOR (Zhu et al. 2021a), we crop the CVACT (Liu and Li 2019) dataset randomly, and define the positive and semi-positive samples.

Evaluation Metrics: For retrieval performance, the conventional top-1, top-K recalls, and hit rate are adopted as the evaluation metrics, following methods (Shi et al. 2020). Hit rate is the percentage of top-1 reference images that cover the query location. For fine-grained localization performance, we report the mean and median distance error over all test samples, following methods (Xia et al. 2022).

Implementation Details. Following VIGOR, the VGG-16 (Simonyan and Zisserman 2014) is adopted as the backbone feature extractor and 8 SAFA blocks (Shi et al. 2019) are used. In a training data unit, each query image corresponds to one positive reference image and one randomly selected from the three sampled semi-positive samples. More details are in the supplementary material.

Ablation Studies

Semantic Ambiguity Modeling. To demonstrate that our framework can properly model the semantic ambiguity of samples, we visualize the predicted localization, uncertainty score, and IOU (between the query and reference images) for each sample. Only the reference images are shown in Fig. 3 for brevity. In (a) and (b), we see that the high uncertainty scores correspond to the low IOU, dynamic foreground, weak discriminating landmark, and large offset prediction error, which verifies that the predicted uncertainty score in the offset regression task can characterize the semantic ambiguity arising from the low overlap rate and big appearance gap between the query and reference images. In (c), the low uncertainty score is accompanied by the high overlap rate, so the similarity in the feature embedding space provides rich information for offset regression. More detailed statistics are provided in the supplementary.

Negative Transfer in Joint Tasks. We model the semantic ambiguity of samples as the uncertainty of query offset, by which the network learns to attenuate the effect of ambiguous samples and thereby mitigates the negative transfer between retrieval and regression tasks. Comparing the second and third rows in Tab. 1, the proposed regression loss $\mathcal{L}'_{reg}$ produces significant performance gain on hit rate (11.2% gains) and mean error ($2.25m$ gains) over $\mathcal{L}_{reg}$, indicating that retrieval performance can be elevated by modeling the ambiguity in regression task, thereby improving regression performance. Fig. 4 displays the learned retrieval and regression performance, characterized by the learning curves of triplet-like and regression losses. Regarding the triplet-like loss, the blue curve obtained by $\mathcal{L}_{trip} + \mathcal{L}_{reg}$ exhibits periodic significant recession, which is due to the adverse impact of the regression task on the retrieval task. In contrast, the red curve using $\mathcal{L}_{trip} + \mathcal{L}'_{reg}$ converges in a more stable and rapid behavior, suggesting that modeling the ambiguity can prevent the network from compromising retrieval performance in favor of regression task. Also, we can observe similar comparisons from the regression losses. These learning curves confirm the effectiveness of modeling semantic ambiguity in mitigating the negative transfer problem between retrieval and regression tasks.

Uncertainty-aware Similarity Metric. To illustrate that our proposed uncertainty-aware similarity metric M_u allows

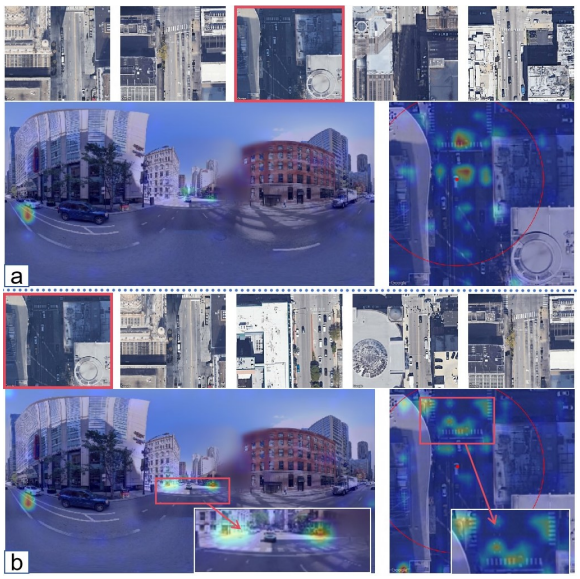

Figure 5: Top-5 retrieval results and feature maps visualization by using the cosine similarity (a) and the proposed uncertainty-aware similarity metric (b). The red boxes denote the correct reference.

the framework to learn at an adaptive pace to deal with the effects of conflicting dominating gradients during the joint training process, we first compare the performances using the proposed ratio loss $\mathcal{L}'_{rat}$ and $\mathcal{L}_{rat}$, where $\mathcal{L}'_{rat}$ and $\mathcal{L}_{rat}$ adopts M_u and cosine similarity as their metric functions respectively. In the $4th$ and $5th$ rows of Tab. 1, $\mathcal{L}'_{rat}$ improves the hit rate to 91.5% (9.7% gains) over the $\mathcal{L}_{rat}$, suggesting the effectiveness of M_u in boosting retrieval performance. A visualization comparison on the feature map is provided in Fig. 5 using Grad-CAM (Selvaraju et al. 2017; Zhu et al. 2021b) to show which regions contribute more to the similarity of the embedding features of two views. As indicated by the rectangles in Fig. 5 (b), the feature maps produced by M_u highlight the zebra crossing on the street, which is the salient landmark consistent across the ground and aerial images. These discriminative features contribute to the correct hit-rate retrieval result. In comparison, the feature maps in Fig. 5 (a) fail to capture the salient landmark and consequently yield incorrect retrieval results. This comparison demonstrates the role of the designed M_u in learning effective feature embeddings for cross view retrieval.

The substantial improvement in the hit rate facilitates a reduction in the mean error of localization ($0.94m$ gains) because more reference images covering the query location could be obtained before the regression task. To further investigate the efficacy of the feature achieved by M_u in addressing the issue of dominating gradients, we compute the ratio of gradient magnitudes generated by the two tasks, as depicted in Fig. 6 (a), following the way in multi-task learning works (Chai et al. 2022; Zhou et al. 2021). With the incorporation of $\mathcal{L}_{rat}$, the blue curves exhibit a decreasing trend during training, indicating that the gradients from

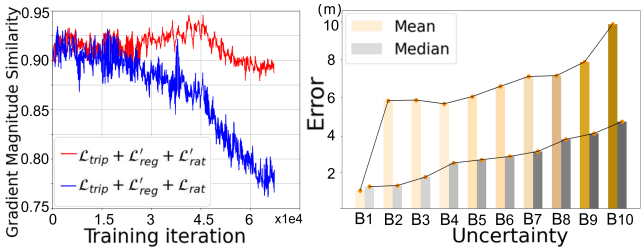

Figure 6: **(a)** Comparison of gradient magnitude similarity. **(b)** The mean error of the regressed query offset across uncertainty intervals ranging from 67.59 to 103.71, divided into 10 equal segments (bins B_i).

the retrieval tasks progressively diminish compared to those from the regression task. This observation suggests that the regression loss induces dominating gradients. In contrast, the red curves obtained with $\mathcal{L}'_{rat}$ approach a value closer to 1.0, signifying a more equitable distribution of gradients between tasks and thus the alleviation of dominating gradients. As a result, $\mathcal{L}'_{rat}$ allows the framework to learn at an adaptive pace, achieving better regression accuracy, as illustrated in Tab. 1. These results demonstrate that the designed M_u is beneficial to handle dominating gradients and improve performance for both tasks.

The Optimal Selection in Top-K Retrievals. During inference, we leverage the modeled semantic ambiguity to select the optimal result from the top-K retrievals. As shown in Tab. 1 (last 2 rows), the top-1 recall improves to 80.6% with this selection, demonstrating the efficacy of the chosen result. This improvement is attributed to the uncertainty's ability to indicate localization quality, enabling the filtering of low-quality predictions. To validate this, we use histogram binning (Schobs et al. 2022) to group test samples by uncertainty values and then compute the mean and median errors of the regressed query offset for each interval (see Fig. 6 (b)). The results show that the bin with the lowest uncertainty corresponds to the smallest error, with a positive correlation between uncertainty and error. Unlike the mean error, the median error consistently increases with rising uncertainty, as it mitigates the bias from outliers with large errors (Fervers et al. 2023). These findings underscore the reliability of our offset uncertainty in indicating localization quality.

Comparisons with State-of-the-Art Methods

VIGOR Dataset: We compare our method with the state-of-the-art methods which consist of *retrieval-only methods*, e.g., TransGeo (Zhu et al. 2022), FRGeo (Zhang and Zhu 2024), Sample4Geo (Deuser et al. 2023), *fine-grained localization only methods*, e.g., MCC (Xia et al. 2022), SliceMatch (Lentsch et al. 2023) and CCVPE (Xia et al. 2023) and retrieval-and-regression localization method VIGOR (Zhu et al. 2021a).

The reported retrieval results for retrieval-only methods are cited from the original papers, using the center point of reference images as localization predictions for practical scenarios. Experiments for only fine-grained localization methods are typically conducted under ideal conditions,

Method	Same-Area								Cross-Area							
	Retrieval (%)				Localization (m)				Retrieval (%)				Localization (m)			
					ideal		pratical						ideal		pratical	
	Top-1	Top-5	Top-10	Hit Rate	mean	median	mean	median	Top-1	Top-5	Top-10	Hit Rate	mean	median	mean	median
TransGeo	61.48	87.54	91.88	73.09	-	-	-	-	18.99	38.24	46.91	21.21	-	-	-	-
FRGeo	71.26	91.38	94.32	82.41	-	-	-	-	37.54	59.58	67.34	40.66	-	-	-	-
Sample4Geo	77.86	95.66	97.21	89.82	-	-	-	-	61.70	83.50	88.00	69.87	-	-	-	-
MCC	-	-	-	-	9.86	4.58	15.67	6.23	-	-	-	-	13.06	6.31	19.45	9.24
SliceMatch	-	-	-	-	5.18	2.58	9.84	4.48	-	-	-	-	5.53	2.55	10.03	4.55
CCVPE	-	-	-	-	3.60	1.36	6.12	2.54	-	-	-	-	4.97	1.68	7.29	3.67
VIGOR	41.07	65.81	74.05	44.71	-	-	19.83	16.59	11.00	23.56	30.76	11.64	-	-	20.09	18.95
Ours	**80.61**	**96.79**	**97.35**	**93.89**	-	-	**4.91**	**1.87**	**63.86**	**85.03**	**89.97**	**72.91**	-	-	**6.28**	**2.95**

Table 2: Comparisons in terms of retrieval and fine-grained localization performance on VIGOR dataset.

with accurate localization priors and no need for global image retrieval; we cite the original results for these ideal scenarios. In practical scenarios, such as GPS denial or large GPS errors, reliable matching reference images are unavailable. For a fair comparison, we use high-performance Sample4Geo to generate initial matching reference images for only fine-grained localization methods, approximating practical conditions. We then reproduce the original codes to predict fine-grained localization results for the retrieved reference images from Sample4Geo. Additionally, the fine-grained localization results for VIGOR in practical scenarios are regression outcomes calculated from the original code.

Tab. 2 shows that: **1)** For the retrieval task, we outperform the second-best method, Sample4Geo, particularly in hit rate, increasing from 89.82% to 93.89%. Unlike Sample4Geo, which uses only positive and negative samples, we design an uncertainty-aware similarity metric that leverages semi-positive samples, enabling more discriminative feature embeddings. This approach improves the accuracy of retrieving reference images covering the query location, enhancing retrieval performance, especially in hit rate. **2)** The performance of fine-grained localization methods significantly degrades in practical scenarios. For example, the mean localization error of SliceMatch increases from $5.18m$ to $9.84m$. These methods assume a reliable reference image fully covering the query, with localization predicted from perfectly aligned images. In practical scenarios, retrieval-based reference images may only partially cover the query or have minimal overlap, severely deteriorating localization performance. **3)** In practical scenarios, our approach outperforms fine-grained (and all other) localization methods, with a localization error of $4.91m$ compared to CCVPE's $6.12m$. First, our improved hit rate over Sample4Geo yields more reference images covering the query location. Second, our framework regresses query offsets for semi-positive samples, enabling accurate localization even with partial coverage. These results show the superiority of our joint framework over fine-grained localization and other methods.

CVACT dataset: In this dataset, each query image is paired with a reference image centered on the query location, which is an ideal condition not reflective of practical scenarios where the query image can be captured anywhere within the reference. To simulate practical scenarios, we resample the reference image into five patches (one central and four corners) each covering the query location. The central patch is treated as a positive reference, while the corner ones are considered as semi-positive, similar to the VIGOR (Zhu et al. 2021a) setting. Tab. 3 reports comparative results where ideal condition ones are cited from the original papers and the practical scenario ones are computed by reimplementing the methods. The top-1 recall of existing methods drops significantly in practical scenarios due to increased semantic ambiguity from semi-positive samples, causing confusion among similar reference images. In practical scenarios, our method achieves the highest top-1 recall (91.02%) and hit rate (94.78%), significantly outperforming other methods. This demonstrates the effectiveness of our method in handling semantic ambiguity raised by semi-positive samples.

Method	Ideal (%)	Pratical (%)	
	Top-1	Top-1	Hit Rate
Shi (Shi et al. 2022)	82.70	76.11	81.41
L2LTR (Yang et al. 2021)	83.14	77.80	83.22
GeoDTR (Zhang et al. 2023a)	85.43	77.92	82.04
TransGeo (Zhu et al. 2022)	84.95	78.08	84.31
UCVGL-base* (Li et al. 2024)	87.89	82.69	88.35
FRGeo (Zhang and Zhu 2024)	90.35	86.74	89.82
Sample4Geo (Deuser et al. 2023)	90.81	87.96	91.63
Ours	-	**91.02**	**94.78**

Table 3: Retrieval accuracy on CVACT-validation dataset.

Conclusion

In this paper, we pioneered the investigation of semantic ambiguity problems arising from semi-positive samples in visual cross view geo-localization and proposed a novel joint retrieval and calibration framework to optimally control the relative task importance automatically. We first modeled the semantic ambiguity during the offset regression process to mitigate negative transfer effects within the joint tasks. Then we introduced an uncertainty-aware similarity metric to propagate uncertainty scores into retrieval task, allowing the model to learn at an adaptive pace to deal with the effects of conflicting dominating gradients during the joint training process. Extensive experiments show the state-of-the-art performances of the proposed modules. We also provide an in-depth analysis of the semantic ambiguity modeling and propagation quantitively and qualitatively. In the future, we will design a more efficient similarity metric with the modeled uncertainty for overall performance improvement.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant (62373293, 62463020, 62403189), in part by Jiangxi Provincial Natural Science Foundation (20242BAB20050), and in part by Ji'an Science and Technology Plan Natural Science Foundation (20244018591). Professor Ajmal Mian is the recipient of an Australian Research Council Future Fellowship Award (project number FT210100268) funded by the Australian Government.

References

Cao, S.-Y.; Zhang, R.; Luo, L.; Yu, B.; Sheng, Z.; Li, J.; and Shen, H.-L. 2023. Recurrent homography estimation using homography-guided image warping and focus transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9833–9842.

Chai, H.; Yin, Z.; Ding, Y.; Liu, L.; Fang, B.; and Liao, Q. 2022. A model-agnostic approach to mitigate gradient interference for multi-task learning. *IEEE Transactions on Cybernetics*.

Chen, Z.; Gupta, A.; Zhou, L.; and Ong, Y.-S. 2022. Scaling multiobjective evolution to large data with minions: a Bayes-informed multitask approach. *IEEE Transactions on Cybernetics*.

Dai, M.; Hu, J.; Zhuang, J.; and Zheng, E. 2021. A transformer-based feature segmentation and region alignment method for uav-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7): 4376–4389.

Dai, M.; Zheng, E.; Feng, Z.; Qi, L.; Zhuang, J.; and Yang, W. 2023. Vision-based UAV self-positioning in low-altitude urban environments. *IEEE Transactions on Image Processing*.

Deuser, F.; Habel, K.; Oswald, N.; and Oswald, N. 2023. Sample4geo: Hard negative sampling for cross-view geolocalisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16847–16856.

Feng, M.; Hou, H.; Zhang, L.; Guo, Y.; Yu, H.; Wang, Y.; and Mian, A. 2023a. Exploring hierarchical spatial layout cues for 3d point cloud based scene graph prediction. *IEEE Transactions on Multimedia*.

Feng, M.; Hou, H.; Zhang, L.; Wu, Z.; Guo, Y.; and Mian, A. 2023b. 3d spatial multimodal knowledge accumulation for scene graph prediction in point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9182–9191.

Feng, M.; Li, Z.; Li, Q.; Zhang, L.; Zhang, X.; Zhu, G.; Zhang, H.; Wang, Y.; and Mian, A. 2021. Free-form description guided 3d visual graph network for object grounding in point cloud. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3722–3731.

Feng, M.; Liu, K.; Zhang, L.; Yu, H.; Wang, Y.; and Mian, A. 2022. Learning from pixel-level noisy label: A new perspective for light field saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1756–1766.

Fervers, F.; Bullinger, S.; Bodensteiner, C.; Arens, M.; and Stiefelhagen, R. 2023. Uncertainty-aware vision-based metric cross-view geolocalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21621–21631.

Hou, H.; Feng, M.; Wu, Z.; Dong, W.; Zhu, Q.; Wang, Y.; and Mian, A. 2024. 3D Object Detection from Point Cloud via Voting Step Diffusion. *IEEE Transactions on Circuits and Systems for Video Technology*.

Hu, W.; Zhang, Y.; Liang, Y.; Han, X.; Yin, Y.; Kruppa, H.; Ng, S.-K.; and Zimmermann, R. 2023. PetalView: Fine-grained Location and Orientation Extraction of Street-view Images via Cross-view Local Search. In *Proceedings of the 31st ACM International Conference on Multimedia*, 56–66.

Kumar, A.; Marks, T. K.; Mou, W.; Feng, C.; and Liu, X. 2019. UGLLI face alignment: Estimating uncertainty with gaussian log-likelihood loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0.

Lentsch, T.; Xia, Z.; Caesar, H.; and Kooij, J. F. 2023. Slicematch: Geometry-guided aggregation for cross-view pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17225–17234.

Li, G.; Qian, M.; Xia, G.-S.; and Xia, G.-S. 2024. Unleashing Unlabeled Data: A Paradigm for Cross-View Geo-Localization. *arXiv preprint arXiv:2403.14198*.

Liu, L.; and Li, H. 2019. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5624–5633.

Lowry, S.; Sünderhauf, N.; Newman, P.; Leonard, J. J.; Cox, D.; Corke, P.; and Milford, M. J. 2015. Visual place recognition: A survey. *IEEE Transactions on Robotics*, 32(1): 1–19.

Lu, F.; Zhang, L.; Lan, X.; Dong, S.; Wang, Y.; and Yuan, C. 2024. Towards Seamless Adaptation of Pre-trained Models for Visual Place Recognition. *arXiv preprint arXiv:2402.14505*.

Schobs, L. A.; Swift, A. J.; Lu, H.; and Lu, H. 2022. Uncertainty estimation for heatmap-based landmark localization. *IEEE Transactions on Medical Imaging*, 42(4): 1021–1034.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Senushkin, D.; Patakin, N.; Kuznetsov, A.; and Konushin, A. 2023. Independent component alignment for multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20083–20093.

Shi, Y.; Liu, L.; Yu, X.; and Li, H. 2019. Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems*, 32.

Shi, Y.; Yu, X.; Liu, L.; Campbell, D.; Koniusz, P.; and Li, H. 2022. Accurate 3-dof camera geo-localization via ground-to-satellite image matching. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 45(3): 2682–2697.

Shi, Y.; Yu, X.; Liu, L.; Zhang, T.; and Li, H. 2020. Optimal feature transport for cross-view image geo-localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11990–11997.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Toker, A.; Zhou, Q.; Maximov, M.; and Leal-Taixé, L. 2021. Coming down to earth: Satellite-to-street view synthesis for geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6488–6497.

Wang, X.; Xu, R.; Cui, Z.; Wan, Z.; and Zhang, Y. 2024. Fine-Grained Cross-View Geo-Localization Using a Correlation-Aware Homography Estimator. *Advances in Neural Information Processing Systems*, 36.

Wei, J.; Yang, Y.; Xu, X.; Zhu, X.; and Shen, H. T. 2021. Universal weighting metric learning for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 6534–6545.

Xia, Z.; Booij, O.; Kooij, J. F.; and Kooij, J. F. 2023. Convolutional cross-view pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xia, Z.; Booij, O.; Manfredi, M.; and Kooij, J. F. 2022. Visual cross-view metric localization with dense uncertainty estimates. In *European Conference on Computer Vision*, 90–106. Springer.

Yang, H.; Lu, X.; Zhu, Y.; and Zhu, Y. 2021. Cross-view geo-localization with layer-to-layer transformer. *Advances in Neural Information Processing Systems*, 34: 29009–29020.

Yu, T.; Kumar, S.; Gupta, A.; Levine, S.; Hausman, K.; and Finn, C. 2020. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33: 5824–5836.

Zhang, Q.; and Zhu, Y. 2024. Aligning Geometric Spatial Layout in Cross-View Geo-Localization via Feature Recombination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7251–7259.

Zhang, X.; Li, X.; Sultani, W.; Zhou, Y.; and Wshah, S. 2023a. Cross-view geo-localization via learning disentangled geometric layout correspondence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3480–3488.

Zhang, Y.; Huang, X.; Zhang, Z.; and Zhang, Z. 2023b. PRISE: Demystifying Deep Lucas-Kanade with Strongly Star-Convex Constraints for Multimodel Image Alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13187–13197.

Zhou, X.; Gao, Y.; Li, C.; and Huang, Z. 2021. A multiple gradient descent design for multi-task learning on edge computing: Multi-objective machine learning approach. *IEEE Transactions on Network Science and Engineering*, 9(1): 121–133.

Zhu, S.; Shah, M.; Chen, C.; and Chen, C. 2022. Transgeo: Transformer is all you need for cross-view image geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1162–1171.

Zhu, S.; Yang, T.; Chen, C.; and Chen, C. 2021a. VIGOR: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3640–3649.

Zhu, S.; Yang, T.; Chen, C.; and Chen, C. 2021b. Visual explanation for deep metric learning. *IEEE Transactions on Image Processing*, 30: 7593–7607.