remote sensing

MDPI

Article

IML-Net: A Framework for Cross-View Geo-Localization with Multi-Domain Remote Sensing Data

Yiming Yan [1,2,3], Mengyuan Wang [1,2], Nan Su [1,2,*], Wei Hou [3], Chunhui Zhao [1,2] and Wenxuan Wang [1,2]

[1] College of Information and Communication Engineering, Harbin Engineering University, Harbin 150009, China; yanyiming@hrbeu.edu.cn (Y.Y.); s321087082@hrbeu.edu.cn (M.W.); zhaochunhui@hrbeu.edu.cn (C.Z.); wangwenxuan@hrbeu.edu.cn (W.W.)
[2] Key Laboratory of Advanced Marine Communication and Information Technology, Ministry of Industry and Information, Harbin 150009, China
[3] Harbin Aerospace Star Data System Science and Technology Co., Ltd., Harbin 150028, China; houwei@spacestar.com.cn
* Correspondence: sunan08@hrbeu.edu.cn

Abstract: Cross-view geolocation is a valuable yet challenging task. In practical applications, the images targeted by cross-view geolocation technology encompass multi-domain remote sensing images, including those from different platforms (e.g., drone cameras and satellites), different perspectives (e.g., nadir and oblique), and different temporal conditions (e.g., various seasons and weather conditions). Based on the characteristics of these images, we have designed an effective framework, Image Reconstruction and Multi-Unit Mutual Learning Net (IML-Net), for accomplishing cross-view geolocation tasks. By incorporating a deconvolutional network into the architecture to reconstruct images, we can better bridge the differences in remote sensing image features across different domains. This enables the mapping of target images from different platforms and perspectives into a shared latent space representation, obtaining more discriminative feature descriptors. The process enhances the robustness of feature extraction for locating targets across a wide range of perspectives. To improve the network's performance, we introduce attention regions learned from different units as augmented data during the training process. For the current cross-view geolocation datasets, the use of large-scale datasets is limited due to high costs and privacy concerns, leading to the prevalent use of simulated data. However, real data allow the network to learn more generalizable features. To make the model more robust and stable, we collected two groups of multi-domain datasets from the Zurich and Harbin regions, incorporating real data into the cross-view geolocation task to construct the ZHcity750 Dataset. Our framework is evaluated on the cross-domain ZHcity750 Dataset, which shows competitive results compared to state-of-the-art methods.

Keywords: geo-localization; multi-domain; IML-Net; ZHcity750

Citation: Yan, Y.; Wang, M.; Su, N.; Hou, W.; Zhao, C.; Wang, W. IML-Net: A Framework for Cross-View Geo-Localization with Multi-Domain Remote Sensing Data. *Remote Sens.* **2024**, *16*, 1249. https://doi.org/10.3390/rs16071249

Academic Editor: Xinghua Li

Received: 20 December 2023
Revised: 10 March 2024
Accepted: 26 March 2024
Published: 31 March 2024

1. Introduction

Cross-view geo-localization technology, as one of the most important and widely applied tasks, involves retrieving the most relevant images of a target from remotely sensed images acquired through different sources. The technology has extensive applications in various fields, such as precision navigation, landmark recognition, trajectory tracking, and unmanned delivery. For instance, when provided with a drone-view image, the system searches for satellite images of the same location from a remote sensing database. Since satellite-view images come with automatic geographic annotations, we can geolocate buildings by accomplishing the matching task between drone-view images and satellite-view images.

In recent years, cross-view geo-localization technology has made significant progress, with its core goal being to learn discriminative and highly generalizable feature descriptors. Cross-view geo-localization works by utilizing deep neural networks (DNNs) with metric

learning to acquire discriminative features [1–5]. Specifically, the network learns a feature space of distances between image pairs that closely match, while increasing the distance between mismatched image pairs [6–8]. Attention mechanisms and directional information are also widely incorporated into network designs [1,2,9]. In the context of remote sensing image matching, where extreme viewpoint changes result in substantial visual appearance alterations, adjacent regions can serve as auxiliary information to enrich discriminative clues for geo-localization [10].

In practical application scenarios, cross-view geo-localization technology is not solely targeted at single remote sensing images but rather encompasses multi-domain remote sensing images. The multi-domain remote sensing images originate from diverse platforms, such as drones and satellites, including images of targets from different perspectives, such as nadir and oblique angles, as illustrated in Figure 1. Additionally, these images may be captured in different seasons, varied weather conditions, and even diverse lighting conditions. Due to the aforementioned characteristics of multi-domain remote sensing images, cross-view geo-localization tasks often pose significant challenges. For instance, in the matching task between drone-view images from different perspectives and satellite-view images, difficulties may arise from incomplete target features due to factors such as changes in perspective and occlusion. To address these issues, we have developed a framework based on real-world application scenarios that can effectively accomplish the task of matching multi-domain remote sensing images.

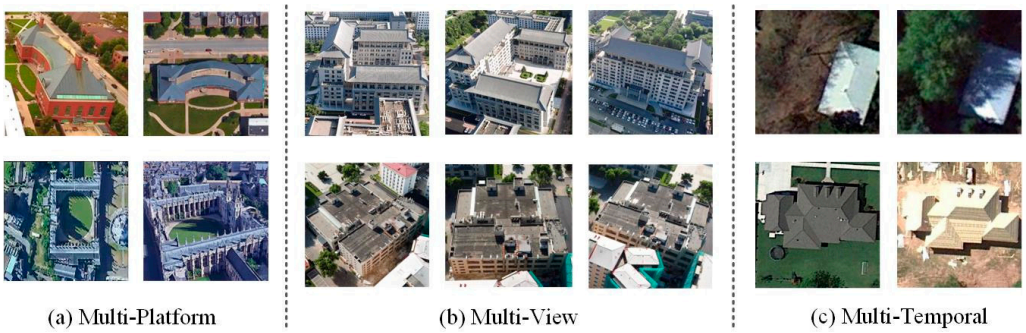

Figure 1. This is a sample collection of multi-domain remote sensing images. (**a**) displays images from various platforms, encompassing satellite views and drone views. The substantial differences between these images pose a challenge for matching. (**b**) showcases drone-view images from different perspectives. Due to the variations in angles, each image may not capture all the information of the target. (**c**) includes satellite-view images captured in different seasons and under varying lighting conditions, which may exhibit significant color variations and instances of occlusion.

Throughout the continuous evolution of cross-view geo-localization technology, datasets tailored for cross-view geo-localization tasks have also emerged. Some early datasets typically provided image pairs, such as those from mobile phone cameras and satellites [2,11]. However, with the rapid development of drone technology, drone-view data are more advantageous in capturing rich information about the target location, as drones flying around the target can provide a comprehensive view with almost no obstacles. Therefore, drones have become the main source for collecting data for cross-view geo-localization tasks. Zheng and others introduced a new dataset, University-1652, to bridge the visual gap between viewpoints [12]. Their work was based on research that increased the training sample size through synthetic data [13–16]. Therefore, the current mainstream datasets for cross-view geo-localization are generated using synthetic drone-view images rather than real images. In fact, there is a significant difference between synthetic and real data. For instance, real drone-view images may face issues such as occlusion due to the large size of the target and a scarcity of training samples due to the high cost of building large-scale real drone datasets. Therefore, widely employed datasets for cross-view geo-localization tasks only include remote sensing images from a single region. Meanwhile, the satellite-view im-

ages in these datasets only consist of single-temporal images, lacking variations in season, weather, and lighting conditions. Such datasets make it challenging to build robust models suitable for real-world scenarios. In real-world applications, the majority of cross-view geolocation tasks involve matching images from different platforms, perspectives, seasons, domains, and styles. In summary, to construct a comprehensive and effective framework for performing cross-view geo-localization tasks in real-world application scenarios, we need a dataset comprising remote sensing images of various cities with different architectural styles. This dataset should include real drone-view images and multi-temporal satellite-view images to learn more comprehensive features of the target.

In this paper, the prominent contributions are as follows:

- A framework named IML-Net, designed to effectively perform the multi-domain remote sensing image-matching task, has been developed. To enhance the stability and robustness of the network, we have created the ZHcity750 Dataset specifically for the task of multi-domain remote sensing image matching. The target areas covered by this dataset encompass two regions with completely different architectural styles. Additionally, the dataset comprises multi-domain remote sensing images, incorporating both real drone-view images and multi-temporal satellite-view images;
- To accomplish the image-matching task for real drone data, the Image Reconstruction Network (IRN) has been proposed, which mainly utilizes the method of reconstructing images to build feature descriptors;
- To enhance the ability of the IRN to construct more discriminative feature descriptors, we have incorporated the Multi-Unit Mutual Learning (MUML) module. This module divides the process into several units and employs the method of mutual learning between different units. This enables the identification of regions with discriminative power in the original image for cropping, using them as augmented data for training.

The rest of this paper is organized as follows. Section 2 reviews and discusses related work. In Section 3, we propose a framework called IML-Net, which can construct cross-domain descriptors through the IRN, and use the MUML to construct augmented training data. Then, in Section 4, we detail the information and process of constructing our multi-domain dataset. Section 5 includes extensive experiments and ablation studies, followed by discussions. In Section 6, we draw conclusions.

2. Related Work

In this section, we will introduce the recent work on cross-view geo-localization and relevant datasets. Next, we will briefly introduce the development of cross-domain feature descriptors and review the methods of learning features in the diverse depths of the layers through neural networks.

2.1. Cross-View Geo-Localization Work and Review of Datasets

The cross-view geo-localization issue is generally viewed as a sub-issue of image matching and retrieval in the field of computer vision. Most previous work is based on two platforms, mobile phone cameras and satellites, to capture datasets for target matching. Oxford5k [17] and Paris6k [18] are the first widely used datasets applied to landmark retrieval. Oxford5k is collected from Flickr and it consists of 5062 images that belong to 11 iconic Oxford buildings, and Paris6k, also from the Flickr platform, differs in that it contains 6412 images of 12 specific Parisian landmarks. One of the earliest works [19] chose to utilize aerial imagery rather than ground views as a geo-reference in order to overcome the shortcomings of ground views that make it difficult to cover a large area of the target region, and constructed a new dataset by selecting both ground-view and aerial-view images from publicly available datasets. The dataset comprises 78,000 pairs of images captured from dual perspectives, 45° bird's eye view and top view. Arandjelovi et al. [20] proposed a method to learn weights based on the image context, allowing the network to focus on regions that contribute positively to the task. Weyand T et al. [21] creatively used geo-grid profiling to transform image matching for a geo-localization issue

into a classification issue; at the same time, it outputs a probability distribution while introducing a dataset containing arbitrary images. This dataset is an arbitrary object taken at the same location with similar characteristics in terms of temporal features. Later, in a similar spirit, Tian et al. [22] borrowed the Google Street View Dataset to build a new dataset and obtained location information by matching the bird's eye view with geographic information to the street view. Differently, they believed that buildings can play an important role in urban localization mandates, so they integrated building detection seamlessly into the entire network architecture and constructed a dataset for urban localization consisting of paired street-view and bird's eye-view images. In addition, the datasets CVUSA [11] and CVACT [4] investigated the challenge of aligning panoramic ground-view images with satellite-view images. The completion of this challenge can be used for user localization when the Global Positioning System (GPS) is not available. The validity of the drone view is confirmed through the introduction of the newly proposed dataset University-1652 [12], which collects 1652 buildings from 72 universities around the world, with data from three views: street view imagery, satellite imagery, and drone imagery, which achieve two new mandates, drone visual localization (Drone→Satellite) and drone visual navigation (Satellite→Drone). Wang T et al. [10] proposed an end-to-end approach to mining environmental information because currently existing approaches usually focus on coarse-grained feature extraction for mining geographic targets at the center of the image while often ignoring the environmental information in the area surrounding the geographic location. Zhang et al. [23] revisit re-ranking and demonstrate that re-ranking can be reformulated as a high-parallelism Graph Neural Network (GNN) function. Tian et al. [24] propose an end-to-end cross-view matching method that integrates a cross-view synthesis module and a geo-localization module, which fully considers the spatial correspondence of drone views and satellite views and the surrounding area information. Dai et al. [25] introduced a simple and efficient transformer-based structure, FSRA, to enhance the model's ability to comprehend text. The structure automatically delineates specific regions based on the weight distribution of the feature map, and these regions can still be delineated and aligned even when there are significant shifts and scale changes in the image. Lin et al. [26] proposed a new framework, RK-Net, to jointly learn discriminative representations and detect saliency key points; the structure contains few learning parameters but significantly improves the performance, and is able to facilitate end-to-end joint learning. Fabian Deuser et al. [27] proposed an orientation-guided drone viewpoint localization training framework based on the estimation of drone image orientations through hierarchical localization to match with satellite images. In the same year, they [28] argued that polar transformations help in matching between different views. However, polar transformations lead to distorted images. So, they proposed contrast learning based on the symmetric InfoNCE Loss. The proposed framework eliminates the need for an aggregation module and avoids further preprocessing steps. It also improves the model's ability to generalize to unknown regions and outperforms the current state-of-the-art results.

Most of the current cross-view geo-localization work is based on the University-1652 [12] dataset, which contains simulated drone-view images, and our real drone images have a series of problems such as incomplete target capture and occlusion due to the angle problem when compared with the simulated images. Meanwhile, the satellite-view images in this dataset only contain one time phase, and there is a lack of satellite images with multiple time phases to assist in different types of tasks. Therefore, we believe that a dataset containing real drone-view images and multi-temporal satellite-view images is needed for cross-view geo-localization, for which the current cross-view geo-localization work still lacks a more effective framework.

2.2. Cross-Domain Feature Descriptors

For the task of cross-view geographic localization, constructing more robust cross-domain feature descriptors is a key challenge. Image feature descriptors, initially designed manually [29–31] and through local feature learning [32,33], have been extensively studied

in the early development of computer vision. With the continuous progress in deep learning, this methodology has been applied to the end-to-end learning of two-dimensional image feature descriptors [34]. In the context of image-matching methods, feature descriptors obtained through self-supervised learning exhibit stronger robustness compared to manually designed descriptors. For example, Zagoruyko et al. [35,36] proposed a Siamese architecture to learn similarity scores between a given set of images. However, these methods incur a relatively high cost as images need to be passed in pairs through networks with the same structure. To address the problem, Simo-Serra et al. [37,38] introduced a method where feature descriptors are learned using the same Siamese architecture but matched using Euclidean distance. The emergence of the highly effective and cost-efficient method has led researchers to replace manually designed descriptors with those obtained through self-supervised learning, and they found that nearest-neighbor queries can be efficiently performed in matching. Compared to network architectures with two images as inputs, a triple network that uses three images as inputs for learning descriptors [39] demonstrates stronger discriminative capability. These triple networks suggest that learning in the absence of Triplet Loss [8,40] results in a better embedding space. Further research has been conducted on enhancing the performance of Triplet Loss [41,42].

Building upon the ideas mentioned above, a cross-dimensional 2D-3D descriptor has been proposed [43], aiming to learn the feature space composed of 2D and 3D descriptors. In recent work, a method was investigated that utilizes Triplet Loss to preserve intra-class similarity and inter-class differences for learning local features, while employing cross-entropy loss to learn globally semantic discriminative features [44].

Our work is rooted in these ideas, with a distinction that we aim to construct a cross-domain feature descriptor. Typically, most efforts in constructing feature descriptors involve taking images from the same domain as inputs, focusing on learning feature spaces containing more useful information for the given task. In contrast, our work concentrates on learning a shared latent space for cross-domain feature descriptors. In addition to employing metric learning, we utilize image reconstruction to learn a more discriminative space.

2.3. Features Learned in the Diverse Depths of the Layers in Neural Networks

Convolutional Neural Networks (CNNs) have demonstrated exceptional achievement in image-matching tasks. Attention learning, as a crucial task in fine-grained recognition [45], helps the model solve a series of problems arising from inherent inter-class similarity and intra-class dissimilarity by capturing more discriminative clues, reducing the importance of semantic information in local regions of target objects. For image-matching tasks, the key is to better capture distinctive feature regions in images and direct the model's attention to these regions. Therefore, it is necessary and effective to understand the diverse levels of semantic information recorded by both deep and superficial layers and to extract information containing distinctiveness through attention learning. Based on this, Zeiler et al. [46] suggested a multi-tiered deconvolutional network to delve into the functions of various feature layers. They discovered that superficial layers acquire low-level details, whereas deep layers acquire high-level semantic information. Jiang et al. [47] pointed out that different layers of CNNs can be used to predict discriminative regions for specific categories. They visualized features extracted from different layers of the network, and the results showed that CNNs gradually shift the model's attention from local details to local regions as the network deepens. Zhang et al. [48] proposed a sequence-diverse network by inserting multiple lightweight sub-networks into the backbone network, enabling information interaction between local regions of fine-grained images, and greatly facilitating the effective learning of different target features. Niu et al. [49], by studying the attention learning process, found that the regions the model focuses on are perceived through attention transfer mechanisms over time. Based on this, they proposed a DNN based on attention transfer mechanisms to find key attention areas and iteratively encode the semantic relevance between the found areas, effectively improving network perfor-

mance. Du et al. [50] focused on which granularity of target regions contributed the most to improving model classification performance and proposed a progressive training strategy to effectively integrate information from different granularities.

Unlike the aforementioned research methods, our approach differs in that we emphasize mutual learning using features extracted from different depths of the network layers. Initially, we use features extracted from layers of varying depths to predict distinct attention areas. The information obtained from these attention areas is then reintroduced into the network for further training, thereby enhancing the model's performance. Through this mutual learning approach utilizing features extracted from different depths of the network layers, we not only assist in training but also increase the quantity of training data.

3. Principle and Methods

In this section, we first provide a detailed discussion of our proposed IML-Net. Subsequently, we illustrate the methodology and principles behind constructing cross-domain descriptors using the IRN. Finally, we present the architecture of the MUML module for building augmented training data.

3.1. IML-Net

The key focus and challenge of cross-view geo-localization technology lie in extracting as complete features of the target building as possible from input drone-view images taken at different angles. Based on this, as illustrated in Figure 2, we proposed the IML-Net to better accomplish cross-view geo-localization tasks.

The framework is primarily composed of the IRN and the MUML module which can assist the model in achieving better cross-domain matching between multi-angle drone images and satellite images. The detailed introductions to these two modules are as follows.

The IRN. The IRN consists of a decoder composed of a 2D encoder and a deconvolutional network. As our research focuses on the multi-domain image-matching problem for geographic localization, our input images can be obtained from drone perspectives at different angles. The images taken at different angles reconstructed by the IRN from the input images are compared to satellite images corresponding to the buildings, and the reconstruction loss is calculated. This process is utilized to construct feature descriptors extracted by the 2D encoder.

Our task is to construct a more robust and distinctive feature descriptor d_i from input images of drone views at different angles. It can be confirmed that building the feature descriptor d_i can be achieved by increasing the similarity between the reconstructed images and the corresponding satellite views of the target buildings. In short, when the reconstructed images closely resemble the features of the original images captured from a vertical perspective., the extracted feature descriptors from the original images can encompass more comprehensive details. The IRN is influenced by the multi-layer deconvolution network proposed by Zeiler et al. [46], which reconstructs the original images through the reverse process of the backbone network, utilizing the transpose of convolution kernels and the reverse feature map calculation. In the deconvolution process, the transpose of the convolution kernels and the reverse feature map calculation are used to return to the previous layer. As for the un-pooling process, we adopt an approximate method, activating the value at the coordinate position of the maximum activation in the pooling process and setting other values to 0. Regarding the activation function, the ReLU function ensures that the activation values of each layer's output are positive. Therefore, for the reverse process, we also use the ReLU function to guarantee that the feature maps of each layer are positive. Our current backbone network employs ResNet50. Due to ResNet50 having residual blocks, it cannot achieve a completely symmetrical reverse process, but the image reconstruction strategy still yields good results.

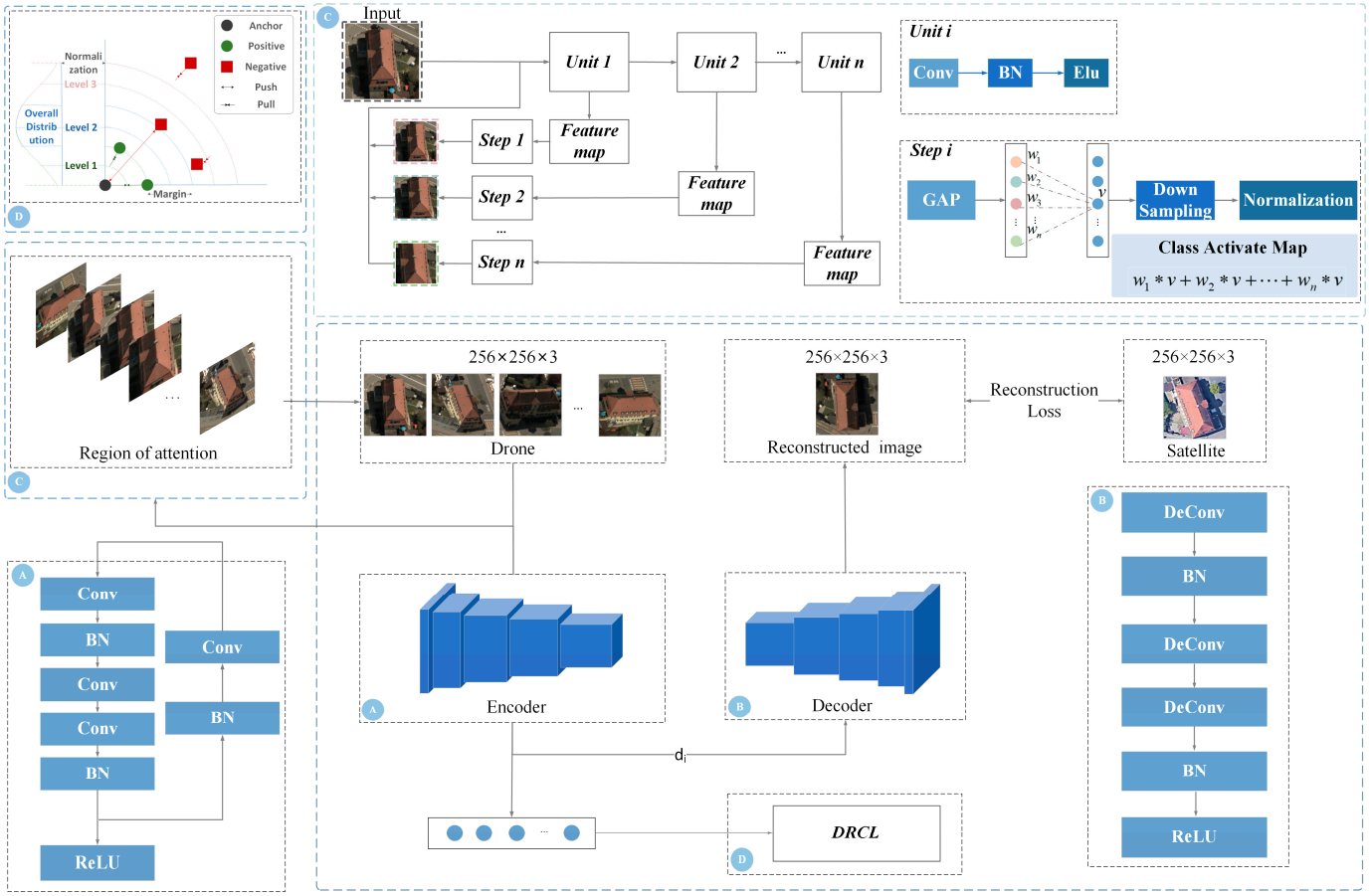

Figure 2. This network model is primarily composed of the IRN and the MUML. The IRN is utilized to construct feature descriptors extracted by the 2D encoder, which consists of a decoder composed of a 2D encoder and a deconvolutional network. The modules A and B in the diagram together constitute the IRN. Due to the insufficient volume of training data to build a more robust model, we employ the MUML to leverage feature information at diverse depths as augmented data during the training process to assist the task. Therefore, we divide the backbone network into four different units to categorize the feature layers of the MUML, as shown in module C in the figure. Module D is the loss function used for computing similarity in the final calculation; we will introduce it in the following sections.

The MUML module. Modern CNNs typically focus on extracting deep-level image features. However, we contend that the shallow-level features extracted by the network are equally important. By facilitating mutual learning between shallow-level and deep-level features, we believe the network can acquire more valuable information. Meanwhile, a small training dataset makes it challenging to build a more robust model, so we employ the MUML to leverage feature information at diverse depths as augmented data during the training process to assist the task. Therefore, we divide the backbone network into four different units to categorize the feature layers of the MUML. The input image passes through different units, generating feature maps of varying sizes. At each step (step *i*), regions of interest are identified in these feature maps, and these regions are cropped from the original image. The cropped images are then reintroduced into the training process as augmented data. Theoretically speaking, through the MUML module, features in the diverse depths of the layers are utilized to learn from each other, constructing attention maps, learning more noteworthy details, and then expanding the training data. This approach serves as a means to enhance the training process and cope with challenges associated with limited training data for building a more robust model.

Through the above two modules, it is possible to effectively extract key details of the target buildings while expanding the training data. This approach can achieve good results.

Our final training loss consists of the reconstruction loss and distance regularization constraint loss:

Reconstruction Loss. The reconstruction loss, defined by mean squared error, represents the loss of the autoencoder network. Specifically, it is the mean squared error between the reconstructed multi-angle drone images D and the corresponding target satellite images S, formulated as follows:

$$\mathcal{L}_{Mse} = \frac{1}{W \times H} \sum_{i=1}^{W \times H} \| S_i - D_i \|^2 \quad (1)$$

In this formula, D_i and S_i represent the i-th pixel in the reconstructed multi-angle drone images and the corresponding target satellite images.

Distance Regularization Constraint Loss (DRCL). To enhance the similarity between target buildings in different domains, i.e., ensure that the multi-angle drone images and their corresponding target satellite images have similar embeddings, most researchers have employed the Triplet Loss function. This type of loss function minimizes the distance between positive samples and the anchor sample while maximizing the distance between negative samples and the anchor sample.

For incorrectly matched noisy samples, when different target buildings are erroneously considered as the same target building, the optimization objective of the Triplet Loss would force the model to learn an infinitely small distance between them. This can lead to overfitting to the noise samples, resulting in a deteriorated final matching performance. To address this issue, we were inspired by [51] to introduce $L2$ regularization and applied distance regularization constraints to optimize the Triplet Loss. We normalize the $L2$ Triplet Loss feature descriptors under the $L2$ norm to lie on a hypersphere with a fixed radius; the aim is to prevent the distance between positive samples and the anchor sample from being minimized, and likewise, to avoid the maximization of the distance between negative samples and the anchor sample, as illustrated in Figure 3:

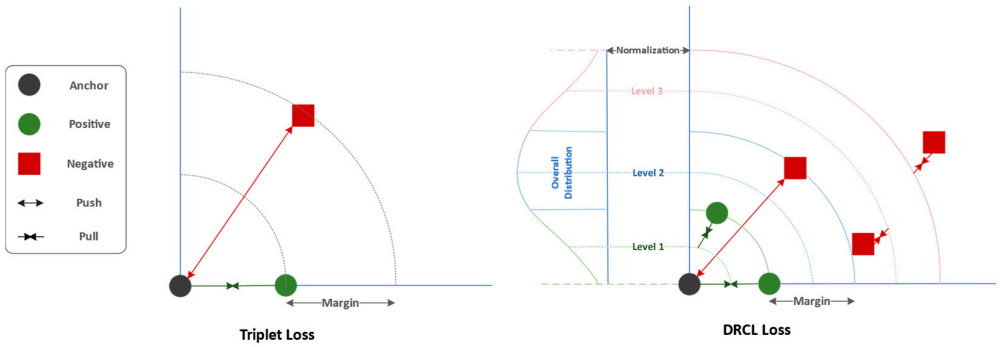

Figure 3. DRCL, compared to Triplet Loss, incorporates an additional $L2$ regularization constraint. This constraint allows the model to learn to shrink the distance between negative samples that are initially too far from the anchor sample to within a certain range. Additionally, it enables the model to learn to increase the distance between positive samples that are initially too close to the anchor sample.

The DRCL is expressed in the following formula:

$$L_{DRCL} = \sum_i^N \left(max\left(F\left(d_a, d_p\right)\right) - min(F(d_a, d_n)) + m, \| f(x_i) \|_2 \right) + \| f(x_i) \|_2 = \alpha, \ \forall i = 1, 2, \cdots, N \quad (2)$$

where m is the margin, F is the distance function, and $\left(d_a, d_p, d_n\right)$ represent the distances corresponding to the anchor sample, positive sample, and the most difficult negative sample, respectively.

Training Loss. Our overall training loss is obtained by multiplying the reconstruction loss and the DRCL by different weights.

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{Mse} + \beta \cdot \mathcal{L}_{DRCL} \tag{3}$$

3.2. IRN for Cross-Domain Descriptors

In the context of multi-angle drone-view images, where the loss of building details is particularly severe, our main objective is to enable the model to learn and construct a more robust feature descriptor. This feature descriptor should contain more detailed information and be able to find the mapping relationship between different views of remote sensing images. Inspired by the deconvolutional network [46], we believe that reconstructing the original input image using a deconvolutional network and calculating the reconstruction loss can help build descriptors that are more favorable for matching. It is important to reconstruct the original image through a deconvolutional network using such feature descriptors. As the reconstructed image features become closer to the target image features, the feature descriptor can contain more comprehensive target information.

The following introduces how the mapping relationship between cross-domain data and feature descriptors is constructed.

Let $\{I_1, I_2, \cdots, I_N\}$ be a set of multi-domain images for the same target location, where $In \in R^{W_n \times H_n \times 3}$ is a color image block of size $W_n \times H_n$, represented in the traditional RGB color space. Each point is represented by its coordinates $(x, y, z) \in R^3$ and RGB color. The goal of learning cross-domain descriptors is to find multiple mappings $\{W_1(.), W_2(.), \cdots, W_n(.), \cdots, W_N(.)\}$, $W_n(.) : R^{W_n \times H_n \times 3} \rightarrow D$, which map the data space of different domains to a shared latent space $D \subseteq R$. Here, R contains common features of different domain data, ensuring that for each set of corresponding relationships between different domains, their mappings are as similar as possible. Mathematically, given the distance function $F(.)$ and descriptors $\{d_1, d_2, \cdots, d_n, \cdots, d_N\}$, where $d_n \in D$, if I and P represent the same underlying geometry,

$$F(d_1, d_2, \cdots, d_n, \cdots, d_N) < m \tag{4}$$

where m is a pre-defined margin.

In addition to constructing the mapping relationship between data in different domains and descriptors, we also aim to learn the inverse functions $f' : D \rightarrow R^{W_n \times H_n \times 3}$ and $g' : D \rightarrow R^{W_t \times H_t \times 3}$. Since these inverse mappings can reconstruct data from descriptors, they prove beneficial in downstream applications, such as visualizing features extracted from diverse depths of the network. In this paper, we utilize the learned cross-domain feature descriptors to reconstruct the original images. Reconstruction is achieved by minimizing the reconstruction loss between the original images and the reconstructed images, serving the downstream task of cross-view geo-localization.

3.3. The Method of MUML for Constructing Augmented Training Data

Modern CNN architectures are typically composed of units [50–52], where a unit refers to a set of layers processing on feature maps with identical spatial dimensions. As depicted in Figure 4, we use units to divide feature layers of diverse depths. The spatial dimensions of the feature maps progressively reduce from the superficial layers to the profound stages. As an illustration, the ResNet50 layers (excluding the fully connected classifier) are organized into four distinct units. When presented with an input image of size 256 × 256 for ResNet50, the spatial dimensions of the output feature maps for layers within the four units decrease from 128 × 128, 64 × 64, 32 × 32, to 16 × 16. After generating discriminative regions through Class Activation Maps (CAMs) [52] on these feature maps of different sizes, attention maps are produced through down-sampling. By normalizing the attention maps of these four units, we can identify and crop regions in the original image that are discriminative. These regions are then used as augmented data in the training process. The specific principles are as follows:

Figure 4. (**a**) shows the drone images from the publicly available dataset provided by the "Benchmark on High Density Aerial Image Matching" project of ISPRS and EuroSDR, as well as the cropped drone-view images; (**b**) contains the satellite images of the target area from Google Maps, along with the preprocessed satellite-view images.

Let C be the backbone network of a neural convolutional network, which can be any CNN developed to date, such as ResNet50, ResNext, etc. C has M layers, where $\{L_1, L_2, \cdots, L_m \cdots, L_M\}$ represents the layers of C from shallow to deep. $\{F_1, F_2, \cdots, F_n \cdots, F_N\}$ are N features in diverse depths of the layer of the network based on M layers. Each feature is composed of features output from a certain layer from L_1 to L_M; for example, F_n is composed of layers from L_1 to L_m, where $1 < m < M$. Features $\{F_1, F_2, \cdots, F_n \cdots, F_N\}$ gradually cover the layers of C from shallow to deep, and the deepest feature F_N covers all layers from L_1 to L_M.

Let $\{M_1, M_2, \cdots, M_n \cdots, M_N\}$ represent the feature maps at an intermediate stage generated by C for features $\{F_1, F_2, \cdots, F_n \cdots, F_N\}$, respectively, in diverse depths of the layer of the network, and $M_n \in \mathcal{R}^{H_n \times W_n \times C_n}$, where H_n, W_n, and C_n represent the height, width, and number of channels of the feature map, respectively. We use a set of functions $\{G_1(\cdot), G_2(\cdot), \cdots, G_n(\cdot), \cdots, G_N(\cdot)\}$ to ensure the reliability of the process of generating feature maps $\{M_1, M_2, \cdots, M_n \cdots, M_N\}$. The functions $G_n(\cdot)$ used to generate feature maps $x\prime(n)$ and $x''(n)$ are defined as follows:

$$x''_n = f^{Elu}\left(f^{bn}\left(f^{conv}_{3\times3\times C_v/2\times C_v}(x'_n)\right)\right) \tag{5}$$

$$x'_n = f^{Elu}\left(f^{bn}\left(f^{conv}_{3\times3\times C_n\times C_v/2}(x_n)\right)\right) \tag{6}$$

where 3×3 refers to the spatial dimension, C_n is the number of input channels, and $C_v/2$ is the number of output channels. $f^{b_n}(\cdot)$ represents the batch sample normalization operation, $f^{Elu}(\cdot)$ represents the Elu operation, and $f^{conv}(\cdot)$ represents the convolution operation with different kernel sizes. For example, $f^{conv}_{3\times3\times C_n\times C_v/2}(\cdot)$ represents a two-dimensional

convolution operation with a kernel size of $3 \times 3 \times C_n \times C_v/2$. The method based on CAM can be used to identify discriminative regions of the image, denoted as $x''(n)$ and $x''(n) \in \mathcal{R}^{H_n \times W_n \times C_v}$.

We define the discriminative region $\phi_n \left(\phi_n \in \mathcal{R}^{H_n \times W_n}\right)$ generated by the CAM method for features F_n in diverse depths of the layer as follows:

$$\phi_n(\alpha, \beta) = \sum_{c=1}^{C_v} p_n x_n''(\alpha, \beta) \tag{7}$$

In this formula, the coordinates (α, β) represent the spatial positions of x'' and ϕ_n, and $p_n \phi_n(\alpha, \beta)$ explains the importance of the spatial position (α, β).

At the same time, we further elaborate on the CAM, which is essentially a linear weighted combination of visual motifs occurring at various spatial positions. These visual patterns are obtained by activating every unit within the intermediate feature map x'', contributing to discriminative regions for image recognition. By up-sampling CAM to obtain regions consistent with the size of input images, we have the capability to understand the most discriminative regions in the image from diverse depths of the feature layers.

Therefore, after obtaining ϕ, we perform down-sampling on ϕ_n using a bilinear sampling kernel to generate an attention map $\widetilde{\phi_n}\left(\widetilde{\phi_n} \in \mathcal{R}^{H_{in} \times W_{in}}\right)$, where H_{in} and W_{in} are the height and width of the input image. Subsequently, we apply min–max normalization to $\widetilde{\phi_n}$, and every spatial component within the normalized attention map $\widetilde{\phi_n^{norm}}$ is defined as follows:

$$\widetilde{\phi_n^{norm}}(\alpha, \beta) = \frac{\widetilde{\phi_n}(\alpha, \beta) - \min\left(\widetilde{\phi_n}\right)}{\max\left(\widetilde{\phi_n}\right) - \min\left(\widetilde{\phi_n}\right)} \tag{8}$$

We obtain the normalized attention map $\widetilde{\phi_n^{norm}}$ to find and crop out the discriminative regions in the features by standardizing the attention map, providing guidance for the matching task. First, we set elements in $\widetilde{\phi_n^{norm}}$ exceeding the threshold $t (t \in [0, 1])$ to 1, and the rest of the elements to 0, generating a mask $\widetilde{\phi_n^{norm}}$. In summary, each spatial element of this mask is given by the following equation:

$$\widetilde{\phi_n^{mask}}(\alpha, \beta) = \begin{cases} 1, & \text{if } \widetilde{\phi_n^{norm}}(\alpha, \beta) - t > 0 \\ 0, & \text{if } \widetilde{\phi_n^{norm}}(\alpha, \beta) - t \leq 0 \end{cases} \tag{9}$$

Similar to the mutual learning mechanism used for fine-grained visual classification [53,54], we also locate a bounding box capable of covering the region of interest highlighted by the mask $\widetilde{\phi_n^{mask}}$. Simultaneously, we find and crop this region from the input image. Subsequently, we resize the cropped region to match the dimensions of the input images through up-sampling. The attention region O_n obtained through this method is treated as additional data introduced during the training process.

Through this process, not only can we assist in training and help the network extract more robust feature descriptors, but it also helps us expand the training data.

4. The Construction of Dataset

In this section, we introduce the ZHcity750 Dataset we constructed, explaining the methods for collecting multi-temporal satellite-view images and capturing multi-angle drone-view images.

We initially selected the publicly available dataset from the "Benchmark on High Density Aerial Image Matching" project by ISPRS and EuroSDR. Specifically, we chose multi-angle drone images captured over the Zurich area in Switzerland. These images were categorized into five perspectives: east, west, south, north, and directly above, based on the drone's shooting orientation. Subsequently, we filtered out buildings with unclear search results and significant changes using Google Maps. From this process, we identified 550 buildings as target locations and encoded their names. The encoding rule involved

arranging the buildings sequentially, treating the drone as a shooting sensor, and considering the five different angles as five frames in a video segment. Finally, we located the encoded buildings in the Google Maps satellite imagery in order of their numbers to obtain corresponding satellite-view images. The selected satellite-view images are the latest available from March 2023 for the target area on Google Maps, and we performed preprocessing, including cropping, on the images. We named this dataset Zurich550, and sample images of this dataset are shown in Figure 5.

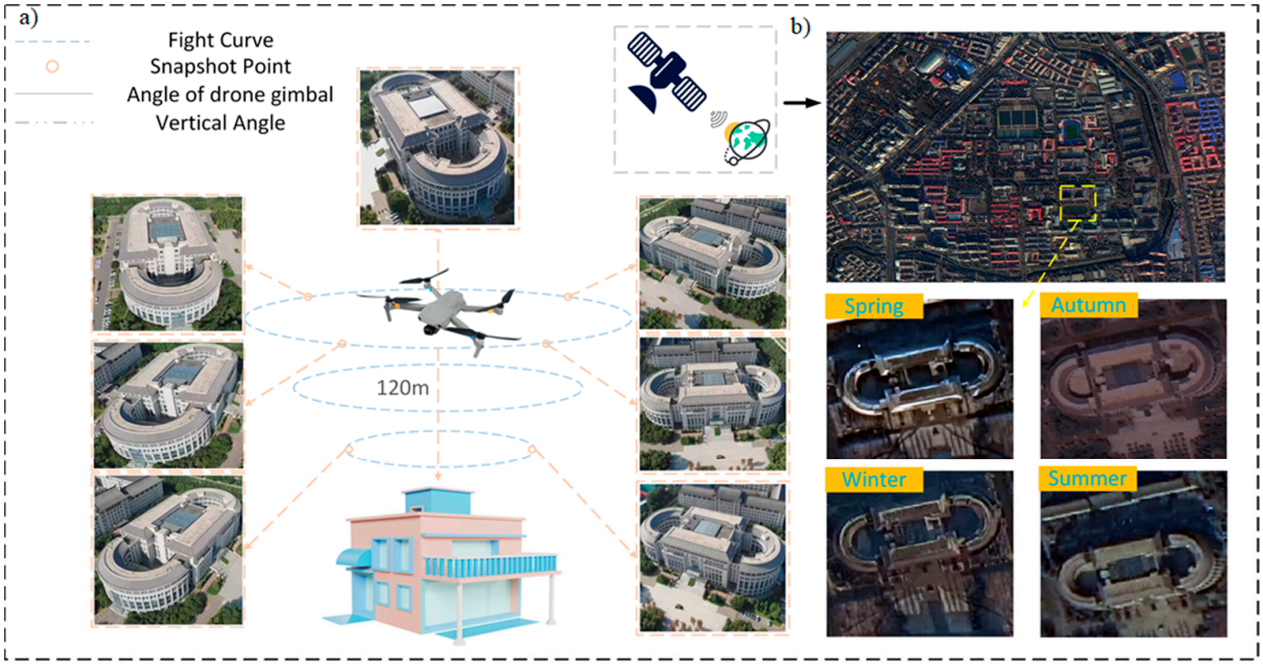

Figure 5. (**a**) shows the drone images from the publicly available dataset provided by the "Benchmark on High Density Aerial Image Matching" project of ISPRS and EuroSDR, as well as the cropped drone-view images; (**b**) contains the satellite images of the target area from Google Maps, along with the preprocessed satellite-view images. The process of creating the Harbin200 dataset. The drone flew around the buildings, capturing three video segments at different altitudes, which were then trimmed. The corresponding target buildings were located in satellite images taken at different times, and four satellite-view images from different seasons were cropped accordingly.

To expand the dataset size and delve deeper into the distinctions between real and simulated data, we chose buildings in Harbin, China, as our target locations. Notably, landmarks were not selected due to two unavoidable issues. First, landmarks often exhibit specific architectural styles that might introduce unforeseen biases. Second, landmarks, being city centers with large crowds, pose safety hazards for drone filming, often prohibiting drone flights in these areas. Considering these challenges, we chose campus and open-area buildings as targets, aligning more closely with real-world practices. Creating the Harbin multi-domain dataset presented significant challenges, involving not only collecting images but also filtering usable information from a vast amount of complex data. Firstly, we selected areas in Harbin city that have not undergone significant changes in recent years as our target areas and chose satellite images with minimal cloud cover and seasonal variations from the Gaofen-2 satellite's extensive archive. Secondly, we encoded the names of buildings in the target area using Google Maps. To develop a model capable of distinguishing subtle differences between buildings, we carefully selected 500 structurally similar buildings. During the preliminary phase of capturing real drone-view images, we conducted field surveys and drone test flights at the selected locations. We discovered that drone flights in Harbin city faced height restrictions. Many of the initially chosen 500 target buildings exhibited an excessive height or oversized footprint, causing excessive

area coverage issues. Therefore, we narrowed down our selection to 200 buildings for drone photography. For the drone-view images of the target buildings, we manually collected drone images of the target buildings (see Figure 4). To achieve scale variation and obtain comprehensive perspectives, we first flew the drone to a designated position, determined the hovering height based on the onsite building height (usually 100–120 m), adjusted the gimbal to shoot at a horizontal angle of 0°, and recorded a 360° flight video at 30 frames per second. Then, we tilted the gimbal down by 30° and 60°, recording flight videos at these angles. Finally, we cropped images from the drone-view videos to obtain seven drone-view images of different scales and perspectives.

Similarly, to ensure a more comprehensive dataset for studying image matching under diverse conditions such as different lighting, weather conditions, and backgrounds, our selection of satellite-view images encompassed all four seasons of the year. The original satellite data were sourced from the Gaofen-2 satellite, equipped with two high-resolution cameras (1 m panchromatic and 4 m multispectral). While the panchromatic images lacked color information, the multispectral images provided spectral information with lower spatial resolution. To obtain high-quality satellite images, we fused panchromatic and multispectral images, and the necessary satellite-view images were derived through cropping and preprocessing the fused satellite images. The dataset for the Harbin area, named Harbin200, was created to include all shooting processes and examples, as illustrated in Figure 4.

In summary, each building in the dataset has an average of 4 satellite-view images and 3 drone-view videos. The satellite images include different seasonal images of Harbin city, while the drone-view videos yield 7 images of varying sizes and perspectives after processing. Additionally, for further research, we can provide satellite-view images of 300 other buildings, each featuring 4 satellite-view images from different times.

Compared to existing datasets (see Table 1), we summarize the capabilities of the ZHcity750 Dataset as follows:

1. Multi-platform. The ZHcity750 Dataset comprises data from two distinct platforms: satellites and real drone footage, as opposed to simulated data;
2. Multi-view. The ZHcity750 Dataset incorporates data from different angles with real drone data images capturing the target building from various perspectives and orientations;
3. Multi-temporal. The ZHcity750 Dataset encompasses data from various temporal phases, featuring multi-temporal data that include satellite-view images captured under different seasons, lighting conditions, and climatic variations.

Table 1. Comparison of the ZHcity750 Dataset with other geolocation datasets. Existing datasets are typically composed of simulated drone data and single-temporal satellite data. In contrast, our dataset focuses on capturing real multi-angle drone images and provides multi-temporal satellite images. Additionally, the table shows the number of images used for training in each dataset, data platforms, target types, evaluation methods, diversity in angles and temporality, and the authenticity of the data.

Datasets		Images for Training	Platform	Type of Target	Evaluation Method	Multi-Angle	Multi-Time Phase	Type of Data
ZHcity750	Zurich550 Harbin200	550 × (5 + 1) 200 × (7 + 4)	Drone, Satellite	Building	Rank@k& mAP	✓ ✓	✗ ✓	Real data
University-1652 [12]		701 × (54 + 16.64 + 1)	Drone, Ground, Satellite	Building	Rank@k&AP	✓	✗	Simulation data
CVUSA [11]		35.5 k × 2	Ground, Satellite	User	Recall@K	✗	✗	Real data
CVACT [4]		35.5 k × 2	Ground, Satellite	User	Recall@K	✗	✗	Real data
Vo et al. [55]		900 k × 2	Ground, Satellite	User	Recall@K	✗	✗	Real data

5. Experiment Results and Discussion

In this section, we first describe the datasets used in the experiment and the evaluation methods, then detail the implementation specifics, followed by providing comparisons with existing techniques and an ablation study.

5.1. Dataset and Evaluation Protocol

Dataset. To validate the performance of the model on real data, this study primarily trains and evaluates our method using the ZHcity750 Dataset we created. Table 2 displays the data distribution for training, testing, and querying in the dataset when addressing various tasks.

Table 2. The data distribution for training, testing, and querying in the ZHcity750 Dataset for the Drone→Satellite and Satellite→Drone tasks.

Dataset		Task					
		Drone→Satellite			Satellite→Drone		
		Train	Test	Query	Train	Test	Query
ZHcity750 Dataset	Zurich550	2000	750	550	2000	1635	2200
	Harbin200	1150	450	200	1150	450	1400

Evaluation Protocol. In our experiments, we employ Rank@1 to evaluate the performance of the model. Rank@1 evaluates the model's ability to correctly identify positive instances, representing the proportion of images accurately matched at the first position in the similarity ranking list. A higher Rank@1 indicates better network performance.

5.2. Implementation Details

We use ResNet50 pretrained on ImageNet as the backbone of our method. To ensure fairness in results during training, we refrain from using data augmentation methods such as cropping and flipping. Each input image is resized to a fixed size of 256×256 pixels. We choose SGD as the optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4}, set the learning rate to 0.01, and employ a batch size of 16. Our model is built in Pytorch, and all the experiments are conducted on an NVIDIA RTX 3090 GPU. Training our network takes approximately 2 h, and we stop after 200 epochs.

5.3. Comparison with State-of-the-Art

Results on Zurich550. As shown in Table 3, we compared our proposed method with other advanced image-matching methods used for geo-localization. The matching accuracy of Rank@1, achieved through our IRN and MUML, reached 77.13% in the Drone→Satellite task and 77.13% in the Satellite→Drone task. The performance of our model on the Zurich550 dataset has exceeded that of other competing methods reported in [10,23,24], and our proposed method has a significant advantage over the best-performing method, with an approximately 3% increase in Rank@1 for both tasks in matching satellite and drone images.

Results on Harbin200. As shown in Table 3, we also trained and tested the model on the Harbin200 dataset and compared it with other advanced methods. Our network achieved a Rank@1 accuracy of 59.25% in the Drone→Satellite task and 63.99% in the Satellite→Drone task. Based on the Harbin200 dataset, compared to the state-of-the-art method FSRA [25], our method showed an accuracy improvement of about 3.2%.

Visualization. As shown in Figures 6 and 7, for additional qualitative evaluation, we displayed the matching results of our network on the ZHcity750 Dataset test set. It is evident that our designed IML-Net can successfully match drone images of different perspectives with satellite images, accurately retrieving the target from various images. In cases where images are incorrectly matched, they exhibit some similar structural patterns

to the query image. When target buildings with nearly identical external conditions appear in the test set, our network architecture encounters issues of target-matching errors.

Table 3. Comparison with the state-of-the-art results reported on the ZHcity750 Dataset.

Training Set	Method	Drone→Satellite		Satellite→Drone	
		Rank@1	AP	Rank@1	AP
Zurich550	RK-Net [25]	66.91	70.12	69.18	73.01
	LPN [10]	72.54	75.99	75.67	79.22
	FSRA [24]	74.83	77.91	78.01	82.35
	IML-Net (Ours)	77.13	79.82	81.19	83.89
Harbin200	RK-Net [25]	51.91	55.94	54.28	58.27
	LPN [10]	53.58	58.67	56.13	60.78
	FSRA [24]	56.62	60.78	60.75	64.32
	IML-Net (Ours)	59.25	63.11	63.99	67.78

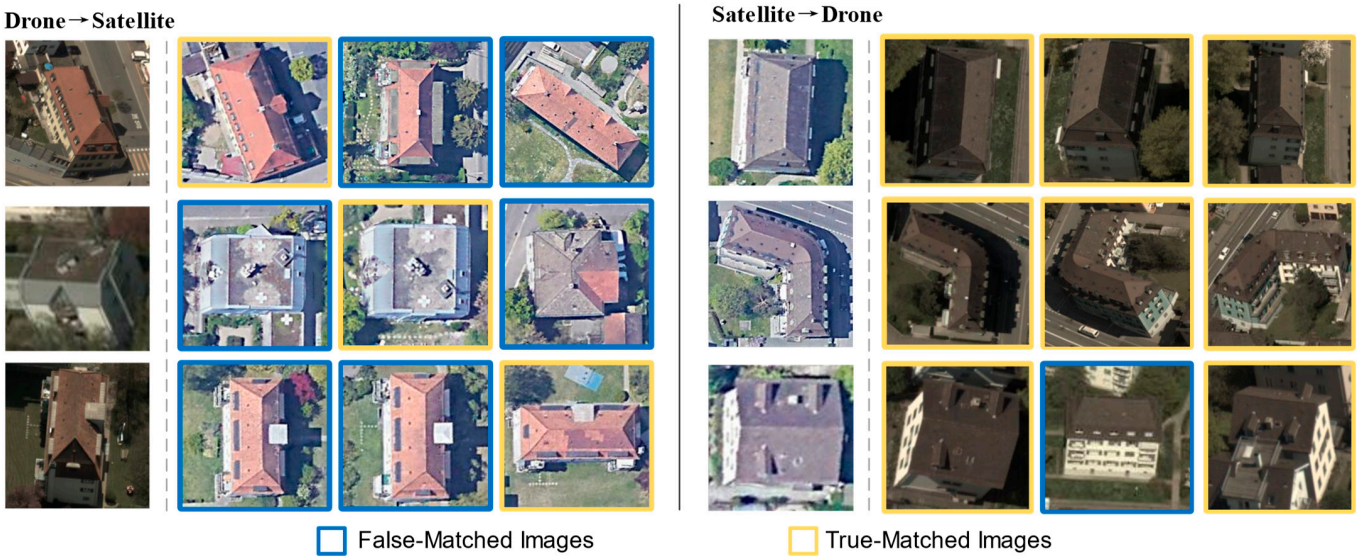

Figure 6. Qualitative image-matching results on the Zurich550 dataset. We present the top three retrieval results for Drone→Satellite (**left**) and Satellite→Drone (**right**). The results are sorted from left to right according to their similarity scores. Images in the yellow boxes are correctly matched images, while those in the blue boxes are incorrectly matched images.

Feature Distribution. To further demonstrate the effectiveness of IML-Net, we visualized the distribution of initial images, features extracted by the baseline, and features extracted by our method on the ZHcity750 Dataset using the t-SNE [56] algorithm. The t-SNE algorithm maps high-dimensional features to a 2D space, and visualization is carried out using the Matplotlib library, as shown in Figure 8. The original data consist of multi-domain data, and due to the feature differences in satellite-view images and drone-view images from different perspectives, their original features are completely separated in the feature space. After introducing the baseline based on ResNet50, the distance between features of the same target building gradually decreases. When replacing the baseline with our IML-Net, the trend of reducing intra-class distance becomes more pronounced. This proves that our method has a positive effect on cross-domain matching.

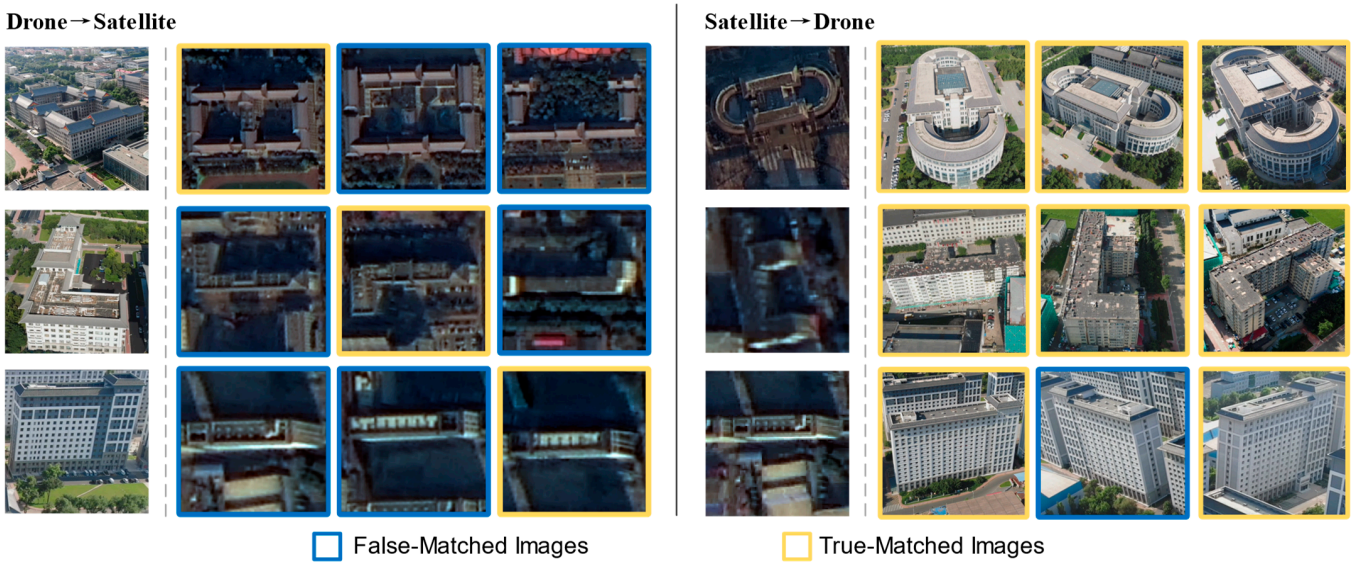

Figure 7. Qualitative image-matching results on the Harbin200 dataset. We present the top three retrieval results for Drone→Satellite (**left**) and Satellite→Drone (**right**). The results are sorted from left to right according to their similarity scores. Images in the yellow boxes are correctly matched images, while those in the blue boxes are incorrectly matched images.

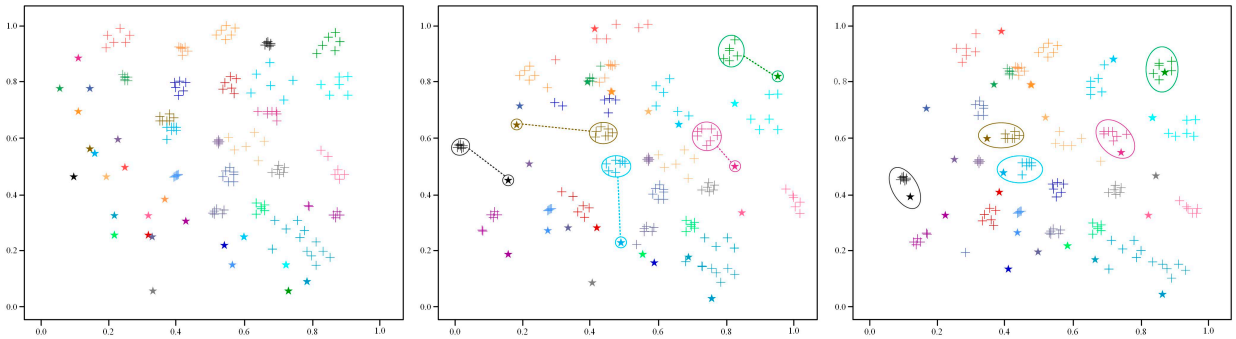

Figure 8. The feature distribution on the ZHcity750 Dataset. We selected 24 identities (IDs) from the dataset, with each ID having 1 satellite-view image and 6 drone-view images chosen to calculate the distance between features. "Stars" and "plus signs" represent satellite-view images and drone-view images, respectively. The circles in the figure contain features of drone-view images and satellite-view images with the same ID. The same color indicates features from the same target, representing intra-class distance, while symbols of different colors represent inter-class distance. (**a**) represents the feature distribution of untrained initial images; (**b**) represents that after the baseline, the intra-class distance gradually decreases; (**c**) shows that, compared to the baseline, the feature distribution of the same class after IML-Net is more conducive to classification.

5.4. Ablation Study

To demonstrate the impact of each module in the network on the matching task, we designed a series of ablation experiments.

Contribution of IRN and MUML. In our evaluation of the individual effectiveness of the IRN and MUML, we integrated each into our baseline separately. As outlined in Table 4, the inclusion of the IRN enhanced the accuracy on Zurich550 by approximately 5%, while the addition of the MUML contributed an extra 2% improvement. Notably, when both the IRN and MUML were integrated into the baseline, the overall accuracy of our network saw

an approximate 7% improvement. This underscores the synergistic impact of the IRN and MUML on our constructed dataset.

Table 4. Ablation study on the effect of the IRN and MUML on ZHcity750.

Baseline	Zurich550	
	Drone→Satellite	Satellite→Drone
	Rank@1	
Baseline	70.49	74.22
Baseline + IRN	75.41	79.35
Baseline + MUML	72.13	77.04
Baseline + IRN + MUML	77.13	81.19

Based on the visualization experiments of the feature distribution in the previous section, we selected a target building from the dataset that had the most similar feature distribution to other targets. We conducted histogram statistics for both intra-class and inter-class distances of this target building. The distances between intra-class and inter-class after the baseline and our architecture are shown in Figure 9. Images of inter-class refer to images of the same target captured from different platforms and perspectives, while images of intra-class refer to images of different targets. After passing through our framework, the distance between images of the same target becomes smaller, while the distance between different targets increases. Additionally, compared to the baseline, the confusion region further decreases. This demonstrates that our IML-Net framework possesses excellent recognition capability.

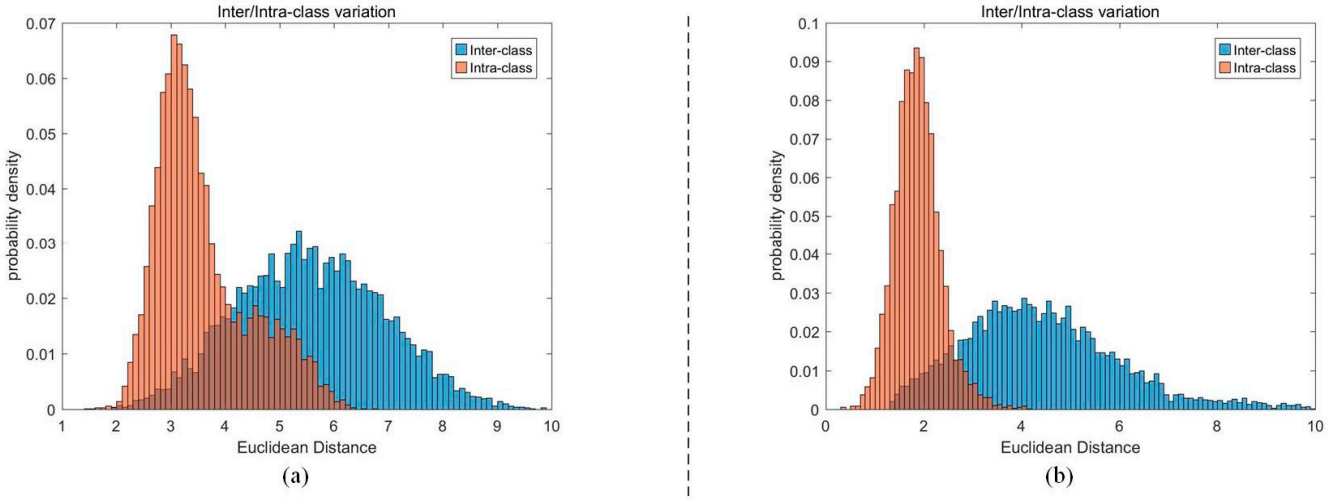

Figure 9. This shows the variations of inter-class and intra-class. (**a**) After the input images go through the baseline, the distance between different targets is very small, and the confusion region is large, making classification challenging. (**b**) After the input images go through IML-Net, the distance between different targets increases, and the confusion region decreases. Compared to the baseline, it exhibits better matching capability.

Impact of the weights of the IRN. As mentioned in Section 3, our training loss is the sum of the IRN's DRCL and reconstruction loss, making the IRN weight one of the most crucial parameters in the experiment. By default, we set the weight to 0.1 to achieve optimal results. When the weight is 0, the reconstruction loss in our training becomes ineffective, and the model is equivalent to the baseline with augmented training data. As shown in Figure 10, an increase in the IRN weight significantly improves the Rank@1 of our model on Zurich550. This suggests that the closer the reconstructed image is to the matching image, the more distinctive the feature descriptor that can be constructed. However, it

is important to note that when the IRN weight exceeds 0.1, the excessively high weight of the reconstruction loss reduces the weight of the DRCL loss. This causes the feature descriptors extracted by the encoder for different angles of drone input images to lack detailed discriminative information, leading to a decline in matching accuracy.

Figure 10. This is the ablation study on the effect of the weights of the IRN on ZHcity750. The blue dashed line represents the accuracy for Drone→Satellite, while the purple dashed line represents the accuracy for Satellite→Drone. The matching accuracy increases with the increase in IRN weight, but when the weight reaches a certain value, Rank@1 actually decreases.

Impact of Input Image Size. Since images contain extremely detailed fine-grained information, compressing the input image size affects the extraction and recognition of this fine-grained information. During the training process, the larger input images will occupy more memory space. To balance these factors, we investigated the impact of input image size on our task. During the experiments, we resized the images without changing the actual size of the targets. As shown in Table 5, on the Zurich550 dataset, there is a noticeable improvement in the accuracy of both Drone→Satellite and Satellite→Drone tasks as the input image size increases from 64 × 64 to 256 × 256. When we further increased the image size to 512 × 512, the matching accuracy began to decline. We hope this experiment helps in selecting the most effective input image size for future experiments, especially when memory constraints are a consideration.

Table 5. Ablation study on the effect of size of inputs on Zhcity750.

Images Size	Zurich550	
	Drone→Satellite	Satellite→Drone
	Rank@1	
64 × 64	69.47	72.33
128 × 128	71.55	75.38
256 × 256	77.13	81.19
384 × 384	76.99	80.96
512 × 512	75.96	80.17

Impact of Backbone. In IML-Net, different backbones result in varying decoder structures, leading to differences in the constructed feature descriptors. The depth of the extracted features also varies, so the network focuses differently on the images, and the images used as augmented data in training can be quite different. To select one or

several backbones that are most helpful for our task from the many widely used options, we experimented with PatchNet [57], VGG16, Res2Next50, and ResNet50 as backbone networks. As shown in Table 6, on the Zurich550 dataset, although ResNet50 has a structure with residual blocks that do not allow for a completely symmetric IRN, it still ensures the highest accuracy among the various backbones. The depths of layers in PatchNet are too shallow to extract discriminative feature descriptors; however, all the remaining networks achieve relatively high matching accuracy. This also demonstrates the generality and effectiveness of the method we proposed.

Table 6. Ablation study on the effect of the backbone on Zhcity750.

Backbone	Zurich550	
	Drone→Satellite	Satellite→Drone
	Rank@1	
PatchNet [57]	19.21	22.33
VGG16	71.55	73.38
Res2Next50	76.46	79.97
ResNet50	77.13	81.19

Impact of Loss Function. To address this question, we conducted a comparison of three different loss functions. As depicted in Table 7, for both Drone→Satellite and Satellite→Drone tasks, the DRCL exhibited a certain degree of improvement compared to the other two loss functions on the Zurich550 dataset. The primary reason for this improvement lies in the fact that the DRCL, as opposed to Triplet Loss, incorporates $L2$ regularization loss to mitigate overfitting of samples and enhance matching accuracy.

Table 7. Ablation study on the effect of loss function on Zhcity750.

Loss Function	Zurich550	
	Drone→Satellite	Satellite→Drone
	Rank@1	
Triplet Loss	76.98	81.04
Soft Margin Triplet Loss	76.91	80.93
DRCL Loss	77.13	80.99

How is the generalization of IML-Net? In our previous experiments, we tended to treat Zurich550 and Harbin200 as two separate datasets for individual training and testing because (1) these two datasets target regions in different countries with significantly different architectural styles, and (2) the satellite-view images included in these datasets are derived from satellite remote sensing images of different resolutions. In real-world scenarios, there may be significant differences in the resolution of images in our query database and the geographical locations of corresponding targets. To investigate whether our network can handle this phenomenon and has a certain level of generalizability, we conducted validation by using either Zurich550 or Harbin200 as the training set and the other as the test set. From Table 8, we observed that IML-Net, compared to the baseline, achieves good performance on both tasks, which also proves that our IRN and MUML contribute to enhancing the model's generalizability. Notably, we observed higher generalizability when the model was trained on Harbin200 and tested on Zurich550 within the same model. Training with Harbin200, which involves matching tasks with real data and increased difficulties, facilitates learning multi-domain feature descriptors crucial for effective matching. Additionally, Zurich550 includes top-down perspective images, which are more similar to satellite-view images, resulting in better performance.

Table 8. Transfer learning from Zurich550 to Harbin200 on IML-Net.

Train Set	Test Set	Baseline		IML-Net (Ours)	
		Satellite→Drone	Drone→Satellite	Satellite→Drone	Drone→Satellite
		Rank@1			
Zurich550	Harbin200	43.3	49.5	48.3	52.7
Harbin200	Zurich550	46.2	52.8	53.6	57.9

6. Conclusions

In this paper, we identify that the primary challenge in current cross-view geo-localization is the absence of a dataset containing real-world data. Therefore, we created a multi-domain dataset comprising true drone-view images from two regions and multi-temporal satellite-view images, with a total size of 5.5 k. Additionally, we propose an effective matching framework for this dataset, constructing cross-domain feature descriptors through the IRN. Specifically, this involves rebuilding the original images based on a deconvolution network strategy, creating features that are more robust and discriminative. To address the issue of limited real drone-view images in our dataset, we employed an MUML module to identify attention regions in the images to expand the training data. On our multi-domain dataset designed for cross-view geo-localization tasks, our approach achieved competitive accuracy compared to three advanced methods: RK-Net [27], LPN [10], and FSRA [26]. Moreover, The IML-Net has better generalizability. The image reconstruction strategy also can be easily integrated into other backbone networks for different tasks. In the future, we plan to research matching multi-temporal satellite-view images within our multi-domain dataset.

Author Contributions: Conceptualization, Y.Y., M.W. and W.W.; methodology, M.W. and Y.Y.; software, data curation and visualization, M.W.; validation, Y.Y., M.W. and W.W.; formal analysis and investigation, W.H., N.S. and C.Z.; resources and funding acquisition, N.S., C.Z. and W.H.; writing—original draft preparation, Y.Y. and M.W.; writing—review and editing, Y.Y.; supervision, C.Z. and N.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset utilized in this research will be made publicly available upon acceptance of the paper. Researchers and interested parties are encouraged to access the dataset for further analysis and verification of the presented findings. Additionally, for access to the codebase employed in this study, please contact the authors via email, and they will be pleased to provide the relevant information.

Acknowledgments: The authors would like to express their sincere appreciation to the ISPRS and EuroSDR for their generous provision of data, particularly for the "Benchmark on High Density Aerial Image Matching" project. The availability of this dataset has played a pivotal role in facilitating our research efforts and contributing to the advancements in the field of high-density aerial image matching.

Conflicts of Interest: The laboratory where author Yiming Yan is affiliated is engaged in joint training programs with the company. Author Wei Hou was employed by the company Harbin Aerospace Star Data System Science and Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Shi, Y.; Liu, L.; Yu, X.; Li, H. Spatial-Aware Feature Aggregation for Cross-View Image Based Geo-Localization. In Proceedings of the 33rd Annual Conference on Neural Information Processing Systems, NeurIPS 2019, Vancouver, BC, Canada, 8–14 December 2019; Volume 32, pp. 10090–10100.
2. Liu, L.; Li, H. Lending Orientation to Neural Networks for Cross-View Geo-Localization. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; Volume 2019, pp. 5617–5626.
3. Shi, Y.; Yu, X.; Liu, L.; Zhang, T.; Li, H. Optimal Feature Transport for Cross-View Image Geo-Localization. In Proceedings of the AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, 7–12 February 2020; pp. 11990–11997.
4. Shi, Y.; Yu, X.; Campbell, D.; Li, H. Where Am I Looking at? Joint Location and Orientation Estimation by Cross-View Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 16–19 June 2020; pp. 4064–4072.
5. Hu, S.; Feng, M.; Nguyen, R.M.H.; Lee, G.H. CVM-Net: Cross-View Matching Network for Image-Based Ground-to-Aerial Geo-Localization. In Proceedings of the 31st Meeting of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7258–7267.
6. Hadsell, R.; Chopra, S.; LeCun, Y. Dimensionality Reduction by Learning an Invariant Mapping. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2006, New York, NY, USA, 17–22 June 2006; Volume 2, pp. 1735–1742.
7. Deng, W.; Zheng, L.; Ye, Q.; Kang, G.; Yang, Y.; Jiao, J. Image-Image Domain Adaptation with Preserved Self-Similarity and Domain-Dissimilarity for Person Re-Identification. In Proceedings of the 31st Meeting of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 994–1003.
8. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A Unified Embedding for Face Recognition and Clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015; pp. 815–823.
9. Fu, Y.; Wang, X.; Wei, Y.; Huang, T.S. STA: Spatial-Temporal Attention for Large-Scale Video-Based Person Re-Identification. In Proceedings of the AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, HI, USA, 27 January–1 February 2019; AAAI Press: 2019; Volume 33, No. 01. pp. 8287–8294.
10. Wang, T.; Zheng, Z.; Yan, C.; Zhang, J.; Sun, Y.; Zheng, B.; Yang, Y. Each Part Matters: Local Patterns Facilitate Cross-View Geo-Localization. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 867–879. [CrossRef]
11. Zhai, M.; Bessinger, Z.; Workman, S.; Jacobs, N. Predicting Ground-Level Scene Layout from Aerial Imagery. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 867–875.
12. Zheng, Z.; Wei, Y.; Yang, Y. University-1652: A Multi-View Multi-Source Benchmark for Drone-Based Geo-Localization. In Proceedings of the 28th ACM International Conference on Multimedia, ACM MM 2020, New York, NY, USA, 12–16 October 2020; pp. 1395–1403.
13. Li, P.; Wei, Y.; Yang, Y. Meta Parsing Networks: Towards Generalized Few-Shot Scene Parsing with Adaptive Metric Learning. In Proceedings of the 28th ACM International Conference on Multimedia, ACM MM 2020, New York, NY, USA, 12–16 October 2020; pp. 64–72.
14. Liu, J.; Ni, B.; Yan, Y.; Zhou, P.; Cheng, S.; Hu, J. Pose Transferrable Person Re-Identification. In Proceedings of the 31st Meeting of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4099–4108.
15. Richter, S.R.; Vineet, V.; Roth, S.; Koltun, V. Playing for Data: Ground Truth from Computer Games. In Proceedings of the 14th European Conference on Computer Vision, ECCV 2016, Amsterdam, The Netherlands, 8–16 October 2016; Volume 9906 LNCS, pp. 102–118.
16. Wu, Z.; Han, X.; Lin, Y.L.; Uzunbas, M.G.; Davis, L.S. DCAN: Dual Channel-Wise Alignment Networks for Unsupervised Scene Adaptation. In Proceedings of the 15th European Conference on Computer Vision, ECCV 2018, Munich, Germany, 8–14 September 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 518–534.
17. Philbin, J.; Chum, O.; Isard, M.; Sivic, J.; Zisserman, A. Object Retrieval with Large Vocabularies and Fast Spatial Matching. In Proceedings of the 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2007, Minneapolis, MN, USA, 17–22 June 2007.
18. Philbin, J.; Chum, O.; Isard, M.; Sivic, J.; Zisserman, A. Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases. In Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, Anchorage, AK, USA, 23–28 June 2008.
19. Lin, T.Y.; Cui, Y.; Belongie, S.; Hays, J. Learning Deep Representations for Ground-to-Aerial Geolocalization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015; pp. 5007–5015.
20. Arandjelovi, R.; Gronat, P.; Torii, A.; Pajdla, T.; Sivic, J. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 5297–5307.
21. Weyand, T.; Kostrikov, I.; Philbin, J. PlaNet—Photo Geolocation with Convolutional Neural Networks. In Proceedings of the 14th European Conference on Computer Vision, ECCV 2016, Amsterdam, The Netherlands, 8–16 October 2016; Volume 9912 LNIP, pp. 37–55.

22. Tian, Y.; Chen, C.; Shah, M. Cross-View Image Matching for Geo-Localization in Urban Environments. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 3608–3616.
23. Zhang, X.; Jiang, M.; Zheng, Z.; Tan, X.; Ding, E.; Yang, Y. Understanding image retrieval re-ranking: A graph neural network perspective. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 16–19 June 2020.
24. Tian, X.; Shao, J.; Ouyang, D.; Shen, H. UAV-Satellite View Synthesis for Cross-View Geo-Localization. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 4804–4815. [CrossRef]
25. Dai, M.; Hu, J.; Zhuang, J.; Zheng, E. A Transformer-Based Feature Segmentation and Region Alignment Method for UAV-View Geo-Localization. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 4376–4389. [CrossRef]
26. Lin, J.; Zheng, Z.; Zhong, Z.; Luo, Z.; Li, S.; Yang, Y.; Sebe, N. Joint Representation Learning and Keypoint Detection for Cross-View Geo-Localization. *IEEE Trans. Image Process.* **2022**, *31*, 3780–3792. [CrossRef] [PubMed]
27. Deuser, F.; Habel, K.; Werner, M.; Oswald, N. Orientation-Guided Contrastive Learning for UAV-View Geo-Localisation. In Proceedings of the 2023 Workshop on UAVs in Multimedia: Capturing the World from a New Perspective, UAVM '23, New York, NY, USA, 23–27 October 2023; pp. 7–11.
28. Deuser, F.; Habel, K.; Oswald, N. Sample4Geo: Hard Negative Sampling for Cross-View Geo-Localisation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, 2–6 October 2023; pp. 16801–16810.
29. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
30. Bay, H.; Tuytelaars, T.; Van Gool, L. SURF: Speeded up Robust Features. In Proceedings of the 9th European Conference on Computer Vision, ECCV 2006, Graz, Austria, 7–13 May 2006; Volume 3951 LNCS, pp. 404–417.
31. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G.R. ORB: An Efficient Alternative to SIFT or SURF. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, 6–13 November 2011; pp. 2564–2571.
32. Analysis, P.; Intelligence, M. Discriminative Learning of Local Image Descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 43–57.
33. Tola, E.; Lepetit, V.; Fua, P. DAISY: An Efficient Dense Descriptor Applied to Wide-Baseline Stereo. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 815–830. [CrossRef] [PubMed]
34. Chopra, S.; Hadsell, R.; Lecun, Y. Learning a Similarity Metric Discriminatively, with Application to Face Verification. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 539–546.
35. Zagoruyko, S.; Komodakis, N. Learning to Compare Image Patches via Convolutional Neural Networks Learning to Compare Image Patches via Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015; pp. 4353–4361.
36. Han, X.; Leung, T.; Jia, Y.; Sukthankar, R.; Berg, A.C. MatchNet: Unifying Feature and Metric Learning for Patch-Based Matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015; pp. 3279–3286.
37. Simo-Serra, E.; Trulls, E.; Ferraz, L.; Kokkinos, I.; Fua, P.; Moreno-Noguer, F. Discriminative Learning of Deep Convolutional Feature Point Descriptors. In Proceedings of the 15th IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, 11–18 December 2015; pp. 118–126.
38. Tian, Y.; Fan, B.; Wu, F. L2-Net: Deep Learning of Discriminative Patch Descriptor in Euclidean Space. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 661–669.
39. Balntas, V.; Lenc, K.; Vedaldi, A.; Mikolajczyk, K. HPatches: A Benchmark and Evaluation of Handcrafted and Learned Local Descriptors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 5173–5182.
40. Hermans, A.; Beyer, L.; Leibe, B. In Defense of the Triplet Loss for Person Re-Identification. *arXiv* **2017**, arXiv:1703.07737v4.
41. Mishchuk, A.; Mishkin, D.; Radenovic, F.; Matas, J. Working Hard to Know Your Neighbor's Margins: Local Descriptor Learning Loss. 2017. In Proceedings of the Advances in Neural Information Processing Systems 30 on Neural Information Processing Systems, NeurIPS 2017, Long Beach, CA, USA, 4–9 December 2017.
42. Keller, M.; Chen, Z.; Maffra, F.; Schmuck, P.; Chli, M. Learning Deep Descriptors with Scale-Aware Triplet Networks Eth Library Learning Deep Descriptors with Scale-Aware Triplet Networks. In Proceedings of the 31st Meeting of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2762–2770.
43. Pham, Q.-H.; Uy, M.A.; Hua, B.-S.; Nguyen, D.T.; Roig, G.; Yeung, S.-K. LCD: Learned Cross-Domain Descriptors for 2D-3D Matching. In Proceedings of the AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, HI, USA, 27 January–1 February 2019; Volume 34, No. 07. pp. 11856–11864.
44. Xiang, X.; Zhang, Y.; Jin, L.; Li, Z.; Tang, J. Sub-Region Localized Hashing for Fine-Grained Image Retrieval. *IEEE Trans. Image Process.* **2022**, *31*, 314–326. [CrossRef] [PubMed]
45. He, J.; Chen, J.; Liu, S.; Kortylewski, A.; Yang, C.; Bai, Y.; Wang, C.; Yuille, A. TransFG: A Transformer Architecture for Fine-Grained Recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, AAAI 2022, Vancouver, BC, Canada, 22 February–1 May 2022; Volume 36, pp. 852–860.
46. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. In Proceedings of the 14th European Conference on Computer Vision, ECCV 2014, Zurich, Switzerland, 6–12 September 2014; Volume 8689 LNIP, pp. 818–822.

47. Jiang, P.T.; Zhang, C.B.; Hou, Q.; Cheng, M.M.; Wei, Y. LayerCAM: Exploring Hierarchical Class Activation Maps. *IEEE Trans. Image Process.* **2021**, *30*, 5875–5888. [CrossRef] [PubMed]
48. Zhang, L.; Huang, S.; Liu, W. Learning Sequentially Diversified Representations for Fine-Grained Categorization. *Pattern Recognit.* **2022**, *121*, 108219. [CrossRef]
49. Niu, Y.; Jiao, Y.; Shi, G. Attention-Shift Based Deep Neural Network for Fine-Grained Visual Categorization. *Pattern Recognit.* **2021**, *116*, 107947. [CrossRef]
50. Du, R.; Chang, D.; Bhunia, A.K.; Xie, J.; Ma, Z.; Song, Y.Z.; Guo, J. Fine-Grained Visual Classification via Progressive Multi-Granularity Training of Jigsaw Patches. In Proceedings of the 16th European Conference on Computer Vision, ECCV 2020, Glasgow, UK, 23–28 August 2020; Volume 12365 LNIP, pp. 153–168.
51. Ranjan, R.; Castillo, C.D.; Chellappa, R. L2-Constrained Softmax Loss for Discriminative Face Verification. *arXiv* **2017**, arXiv:1703.09507v3.
52. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2921–2929.
53. Liu, D.; Zhao, L.; Wang, Y.; Kato, J. Learn from Each Other to Classify Better: Cross-Layer Mutual Attention Learning for Fine-Grained Visual Classification. *Pattern Recognit.* **2023**, *140*, 109550. [CrossRef]
54. Li, Y.; Chen, L.; Li, W.; Wang, N. Few-Shot Fine-Grained Classification with Rotation-Invariant Feature Map Complementary Reconstruction Network. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–12. [CrossRef]
55. Vo, N.N.; Hays, J. Localizing and Orienting Street Views Using Overhead Imagery. In Proceedings of the 14th European Conference on Computer Vision, ECCV 2016, Amsterdam, The Netherlands, 8–16 October 2016; Volume 9905 LNCS, pp. 494–509.
56. Laurens, V.D.M.; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
57. Wang, C.Y.; Lu, Y.D.; Yang, S.T.; Lai, S.H. PatchNet: A Simple Face Anti-Spoofing Framework via Fine-Grained Patch Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, 19–23 June 2022; pp. 20281–20290.