Article

Co-Visual Pattern-Augmented Generative Transformer Learning for Automobile Geo-Localization

Jianwei Zhao [1,2], **Qiang Zhai** [1,2,*], **Pengbo Zhao** [3], **Rui Huang** [1,2] and **Hong Cheng** [1,2]

1   School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
2   Center for Robotics, University of Electronic Science and Technology of China, Chengdu 611731, China
3   McCormick School of Engineering, Northwestern University, Evanston, IL 60611, USA
*   Correspondence: qzhai@std.uestc.edu.cn

Abstract: Geolocation is a fundamental component of route planning and navigation for unmanned vehicles, but GNSS-based geolocation fails under denial-of-service conditions. Cross-view geo-localization (CVGL), which aims to estimate the geographic location of the ground-level camera by matching against enormous geo-tagged aerial (e.g., satellite) images, has received a lot of attention but remains extremely challenging due to the drastic appearance differences across aerial–ground views. In existing methods, global representations of different views are extracted primarily using Siamese-like architectures, but their interactive benefits are seldom taken into account. In this paper, we present a novel approach using cross-view knowledge generative techniques in combination with transformers, namely mutual generative transformer learning (MGTL), for CVGL. Specifically, by taking the initial representations produced by the backbone network, MGTL develops two separate generative sub-modules—one for aerial-aware knowledge generation from ground-view semantics and vice versa—and fully exploits the entirely mutual benefits through the attention mechanism. Moreover, to better capture the co-visual relationships between aerial and ground views, we introduce a cascaded attention masking algorithm to further boost accuracy. Extensive experiments on challenging public benchmarks, i.e., CVACT and CVUSA, demonstrate the effectiveness of the proposed method, which sets new records compared with the existing state-of-the-art models. Our code will be available upon acceptance.

Keywords: cross-view geo-localization; vision transformer; deep representation learning

1. Introduction

Geolocation identification of automobiles has been a topic of growing interest in recent years due to its potential applications in navigation and route planning for intelligent vehicles [1–7]. Conventionally, obtaining the geographic location of a vehicle through Global Navigation Satellite Systems (GNSS) has been a convenient and cost-effective method. However, GNSS signals are prone to being unreliable or unavailable due to the presence of dense high-rise obstacles, network failures, etc. For example, scenarios such as dense primordial forests and crowded buildings are shown in Figure 1. Fortunately, current satellite images can cover most outdoor scenarios where automobiles are involved and are easily collected offline in advance through open services like Google Maps. To overcome this limitation, the use of registered ground–satellite image retrieval for geographic location estimation has gained increasing attention [8–14]. This method involves the comparison of visual data obtained from the vehicle with geo-tagged references stored in a database, resulting in the estimation of the geographic location that is aligned with the closest reference. This pipeline is schematically illustrated in Figure 1.

Typically, geolocation involves the collection of perspectives from sites previously visited by vehicles. Upon subsequent revisits, these views can be compared with similar

scene content, constituting a loop closure detection process. In the event of satellite signal failure, the agent is required to determine its position by analyzing the contextual scene. Such methodologies are designed to mitigate ambiguity by exploring and encoding contextual information and deep semantics. The images are encoded by identical or Siamese-like backbone networks, followed by nearest-neighbor matching. Thus, the geolocation task is akin to image retrieval, albeit with a primary focus on capturing and leveraging geometric and structural information of the environmental features that constitute the scene. Such information may include, but is not limited to, edge and corner features, shapes, and their relative positions, all of which are fundamental to effective geolocation. Consequently, geolocation requires a more nuanced understanding of the scene content than traditional image retrieval, as it must incorporate this rich geometric and structural information into its matching algorithm to achieve accurate results. Overfeat [15] was a pioneering deep learning-based study in the field and inspired a series of improvements [8,16–19]. To construct a reference data set with GPS information, these approaches examine the ground-to-ground matching procedure for localization by gathering views at diverse locations at different times, seasons, and weather conditions, as exemplified by Google Street View, a widely-used application. During the localization phase, views with unknown locations are matched with reference sets to estimate their locations. Despite their effectiveness, these methods are labor-intensive and cannot locate places that are not in the reference dataset. Therefore, researchers are striving to establish interconnectivity between satellite views and ground views by extracting the intrinsic similarities between the two view types, namely cross-view geo-localization (CVGL), which increases the generalization performance of the location model. Owing to the dissimilar imaging perspective between satellite and ground views, the appearance of content varies significantly, posing a substantial challenge in achieving cross-view localization. Nonetheless, researchers have made remarkable strides in devising Siamese-like networks that contain two distinct branches responsible for encoding each view independently [14,20–29]. While the relationship between different views provides a significant impetus for cross-view localization, several challenges persist. **First**, semantic consistency between views is not fully leveraged. Current methods typically utilize Siamese-like networks for independent encoding of cross-view views but often neglect the high-order consistency semantics of view content, which is essential for matching ground and satellite images. **Second**, co-visual relationships between views are not explicitly accounted for. The perspective disparities between ground and satellite views limit co-visual relationships exploring, with the latter typically encompassing a more extensive scope; thus, using the whole image for coding would yield suboptimal accuracy. **Third**, deep contextual semantic mining is not yet sufficient. As the interaction between views remains unconsidered, the existing methods fail to fully explore contextual semantics.

To address the above deficiencies, we present a novel mutual generative transformer learning (MGTL) for the CVGL task. We first revisit the attention learning strategy and propose a novel cascaded attention masking algorithm to create network reasoning for the co-visual patterns between ground and satellite views. Then, two symmetrical generative sub-modules, i.e., Ground-to-Satellite (G2S) and Satellite-to-Ground (S2G), are thoughtfully designed to generate the simulated cross-view knowledge and to capitalize on the mutual benefits across views. Specifically, S2G takes the aerial semantics and skillfully simulates the ground-aware knowledge, and vice versa. Subsequently, the view-specific simulated knowledge is applied to strengthen the current view features via attention learning, and all the sub-components work in concert within a transformer-based framework to accomplish the CVGL task. The experimental results for several challenging public benchmarks unequivocally establish the superiority of our proposal. The contributions of the proposed MGTL can be summarized as follows:

- **A novel cross-view knowledge-guided learning approach for CVGL.** To the best of our knowledge, the MGTL is the first attempt to build mutual interactions between ground-level and aerial-level patterns in the CVGL community. Unlike existing

transformer-based CVGL models that only perform self-attentive reasoning in the respective view, our proposed MGTL produces cross-knowledge information to achieve more representative high-order features.

- **Cascaded attention-guided masking to exploit the co-visual patterns.** Instead of treating patterns in aerial and ground views equally, we developed an attention-guided exploration algorithm to create network reasoning based on the co-visual patterns, which further improves performance.
- **State-of-the-art localization accuracy on widely-used benchmarks.** The proposed MGTL outperforms existing deep models on various datasets, i.e., *CVUSA* [30] and *CVACT* [31], as shown in Figure 2.
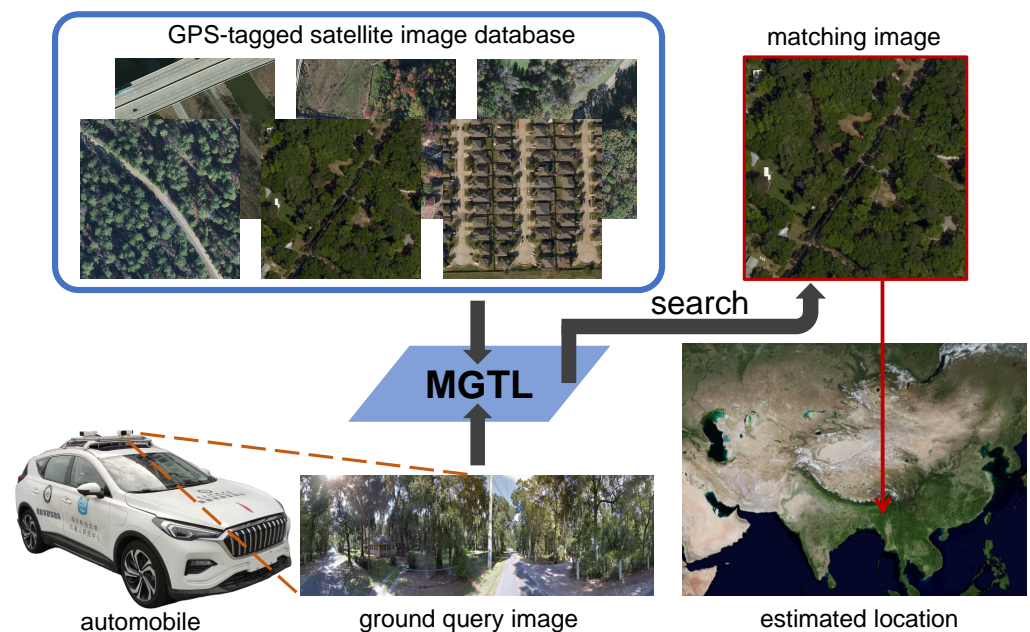
Figure 1. Schematic diagram of image matching-based cross-view geo-localization.
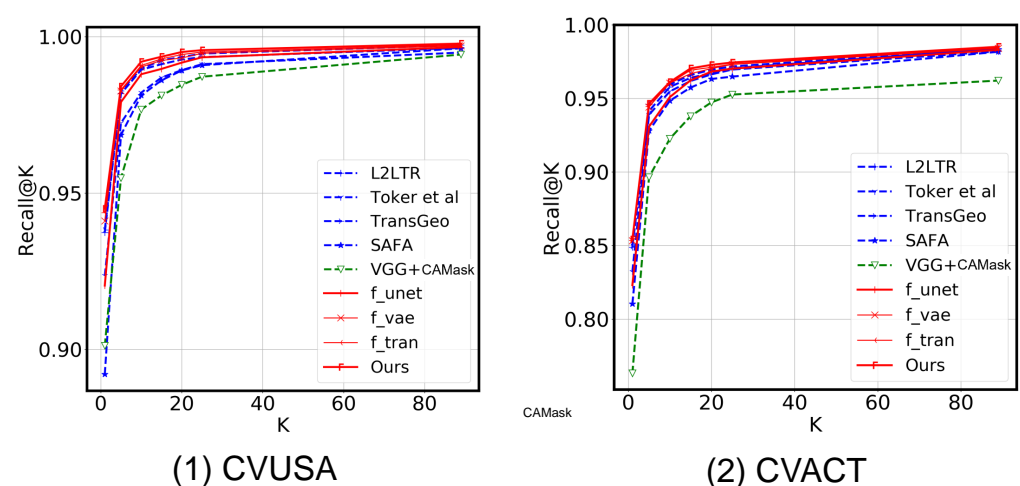
Figure 2. The proposed MGTL outperforms existing approaches.

2. Related Work

Visual Place Recognition: Visual place recognition (VPR) involves matching ground-to-ground images and is a crucial aspect of vision-based navigation and localization, particularly in the field of autonomous driving. The problem tackled by VPR is to determine the current camera's location in the existing image database, and the task becomes

challenging due to factors such as seasonal variations and dynamic visual object changes. A primary approach to overcoming this challenge involves extracting high-level feature descriptors from input images and comparing them based on distance. To improve VPR performance, Arandjelovic et al. [8] modified the traditional non-differentiable operation in vector of locally-aggregated descriptors (VLAD) and incorporated it into CNN-based networks to develop an end-to-end trainable VLAD descriptor named NetVLAD. Following the success of NetVLAD, several variants [17,18] have been proposed. To exploit the multi-scale information, spatial pyramid-enhanced NetVLAD (SPE-NetVLAD) [18] integrated multi-scale features in the training phase by cascading encoding features with varying scales in the final convolutional layer of NetVLAD to improve the performance of VPR. Multi-resolution NetVLAD (MultiRes-NetVLAD) [17] utilized low-resolution image pyramid coding and presented a multi-resolution residual aggregation scheme to enhance the NetVLAD learning feature representation ability. In addition, to address the issue of seasonal and time-of-day variations, Latif et al. [19] approached the VPR problem as a region translation task. A pair of coupled generative adversarial networks (GANs) was utilized to generate the appearance of one domain from another without requiring image-to-image correspondences across the domains. These classical solutions in the field of VPR offer valuable insights for addressing the challenge of cross-view image matching.

Cross-View Geo-Localization: Current cross-view geo-localization (CVGL) pipelines utilize a Siamese-like neural network to extract feature representations from each view, followed by the definition of a metric that places the embedding features of cross-view images in close proximity based on their GPS coordinates. The primary obstacle in CVGL tasks is the significant appearance gap between ground and aerial views caused by changes in viewpoint [32]. Satellite-view images are typically composed of satellite images captured by specialized panchromatic and multispectral cameras on board satellites, whereas ground-view images consist of panoramic images taken using handheld or vehicular optical cameras. These two images have different imaging principles and shooting angles, leading to stark differences in image appearance, such as the representation of visual objects and their spatial layout. This problem is further exacerbated by the large time intervals between acquisition of images. Prior work has mainly addressed this issue by focusing on extracting viewpoint-invariant features [28,29,33] or applying viewpoint transformation [34–36]. The former involves designing effective network architectures that can extract invariant features across views. Workman et al. [9] proposed a convolutional neural network (CNN) to learn a joint semantic feature representation for aerial and ground-level imagery, while Lin et al. [37] introduced a Siamese-like network followed by Euclidean distance calculation to measure cross-view feature representation similarity. More recently, Hu et al. [11] utilized NetVLAD to encode global descriptors and a Siamese-like CNN-based network to extract local feature descriptors for more robust representation learning. Sun et al. [38] further presented a pure convolutional network equipped with capsule layers to model the spatial feature hierarchies. In contrast, to address the imagery geometric gap caused by viewpoint differences, Shi et al. [23,39] used polar transform and attention mechanisms to pre-process satellite imagery, which has been shown to be highly effective. Recently, Yang et al. [14] and Zhu et al. [13] proposed transformer-based methods, leveraging self-attention mechanisms to model global dependencies. Zhu et al. [13] introduced a novel attention-based masking mechanism to remove redundant areas in satellite images, reducing interference in matching performance. The latter approach involves exploring ways to synthesize realistic cross-domain imagery using viewpoint transformation. Ren et al. [40] proposed a cascaded cross MLP-mixer GAN (CrossMLP) module to extract latent mapping cues between cross-view imagery, while Toker et al. [41] developed a GAN-based multi-task architecture to synthesize realistic street views from satellite images. However, existing methods lack mutual learning across views and fail to consider the inter-dependencies between latent features in different network branches. In this paper, we propose a novel approach that integrates cross-view knowledge generative tactics into the transformer architecture, referred

to as mutual generative transformer learning. This approach leverages mutual learning across different views to improve feature representation ability and retrieval performance.

Vision Transformer: The transformer [42] has gained widespread use in the field of natural language processing (NLP) due to its excellent global modeling ability and self-attention mechanism, as demonstrated by its superior properties [42]. The self-attention mechanism is based on the calculation of dot product similarity by query and key, which are then multiplied with value, where query, key, and value represent different embedding spaces computed by the input feature sequence. Dosovitskiy et al. [43] introduced the Vision Transformer (ViT), which is a modified version of the standard transformer that takes the embedding sequences of image patches with $k \times k$ resolution as input [43]. Unlike the standard transformer in NLP, ViT discards the locality assumption and requires less vision-specific sensing bias, dominating in classification [44–46], semantic segmentation [47–49], object detection results [50–52], super-resolution restoration [53,54], depth estimation [55,56], etc. Chen et al. [44] proposed a multi-headed and multi-tailed shared backbone structure to cope with different vision tasks. Lanchantin et al. [46] proposed the classification transformer (C-Trans) network to complete a generic multi-label image classification task. Segmentation transformer (Segmenter) [47] defined the semantic segmentation task as a sequence-to-sequence problem and employed the transformer architecture. Zheng et al. [49] incorporated different decoders into ViT to tackle segmentation tasks. Detection transformer (DETR) [50] employed a transformer-based approach and treated object detection as a set prediction problem. Misra et al. [51] added non-parametric queries and Fourier positional embeddings to the traditional transformer to suit the 3D object detection task. Zamir et al. [54] modified several key designs in a multi-head attention and feed-forward network so that they can capture long-range pixel interactions while still being suitable for high-resolution images. Liang et al. [53] introduced the ViT into light field image super-resolution restore tasks. Li et al. [55] utilized dense pixel matching with location information and attention mechanisms in the transformer to take the place of customer construction widely used for depth estimation. Ding et al. [56] designed a novel end-to-end deep neural network based on a feature matching transformer (FMT). In addition to RGB image fields, current works have also scrutinized the application of transformers in hyperspectral images (HSI) [57–60] and achieved superior results. He et al. [57] introduced a new spatial–spectral transformer (SST) classification framework comprising an improved dense transformer layer for HSI classification. Sun et al. [59] improved a spectral—spatial feature tokenization transformer (SSFTT) method to capture spectral-–spatial features and high-level semantic features. Multispectral fusion transformer network (MFTNet) [60] was designed as a novel feature fusion tactic to generate robust cross-spectral fusion features. Researchers have proposed a series of variants to improve the general ability of ViT. These variants contain substantial skillful tactics such as enhanced locality, improved self-attention algorithms, and structural redesign [61–65]. To introduce the locality principle in the transformer, Chu et al. [61] proposed the conditional positional vision transformer (CPVT), which uses a conditional positional encoding scheme consisting of a 2D CNN to realize translation invariance. Positional embeddings are generated based on the local relationship of the restricted tokens, which encode the relative location information of tokens implicitly [61]. Locality vision transformer (LocalViT) [62] is inspired by the comparison between feed-forward networks (FFN) and reverse residual blocks, and depth-wise convolutional is applied to FFN to add locality to the vision transformer [62]. Cross-scale attention transformer (Cross-Former) [66] presented multi-scale feature representation learning tactics in combination with a vision transformer. Cross-attention multi-scale vision transformer (CrossViT) [63] proposed a two-branch transformer to process tokens generated by patches of different sizes and then fused these tokens multiple times to achieve mutual complementation of semantic information by applying cross-attention interaction [63]. Liu et al. [64] proposed a hierarchical vision transformer using shift windows (swin-transformer), using a shift-window-based module to replace the traditional multi-head self-attention. The framework allows for cross-window connections and promotes the flexibility of modeling at different

scales. Considering the transformer's powerful global modeling ability and successful application in visual works, we designed a transformer-based network to further explore its potential in the cross-view geo-localization task.

3. Methods

3.1. Problem Formulation

Let the cross-view geo-localization (CVGL) model be indicated as the function $\mathcal{F}_{\Theta}$ parameterized by weights Θ, which takes an image pair consisting of a ground-view image $\mathbf{I}_G$ and a satellite-view image $\mathbf{I}_S$ as input and produces their corresponding representations $\mathbf{F}_G$ and $\mathbf{F}_S$. Our goal is to learn Θ from the labeled training triplets $\{\mathbf{I}_G^i, \mathbf{I}_{SP}^i, \mathbf{I}_{SN}^i\}_{i=1}^N$ to make $\mathbf{F}_G$ and $\mathbf{F}_S$ closer while their corresponding cross-view images are matching, where $\mathbf{I}_G^i$ is the ground-view image and $\mathbf{I}_{SP}^i$ and $\mathbf{I}_{SN}^i$ are the positive and negative samples relative to $\mathbf{I}_G^i$, respectively. The process can be formulated as follows:

$$\mathbf{F}_G, \mathbf{F}_S = \mathcal{F}_{\Theta}(\mathbf{I}_G; \mathbf{I}_S), \tag{1}$$
$$||\mathbf{I}_G^i - \mathbf{I}_{SP}^i||^2 + \alpha < ||\mathbf{I}_G^i - \mathbf{I}_{SN}^i||^2$$

where α is the margin in the triplet loss.

3.2. View-Independent Feature Extractor ($f_{\mathtt{VIFE}}$)

3.2.1. Overview

We retained the initial 13 convolutional layers in VGG16 [67] and split them into 5 stages according to spatial resolutions in order to extract high-order features from input images. Then, we designed a cascaded attention-masking (CAMask) algorithm for learning fine-grained co-visual relationships by cascading multi-branch convolutional modules. Figure 3 illustrates the overview of our proposed mutual generative transformer learning (MGTL). As mentioned above, $f_{\mathtt{VIFE}}$ takes an image pair <$\mathbf{I}_G$, $\mathbf{I}_S$> as input and produces two view-specified semantic representations <$\mathbf{F}_G^{'}$, $\mathbf{F}_S^{'}$> and corresponding spatial attention masks <$\mathbf{M}_G$, $\mathbf{M}_S$>, following Equations (2) and (3). Additionally, we have listed the main abbreviations in Table 1 for ease of reference.

Table 1. List of **Abbreviations.**

Abbreviation	Explanation	Abbreviation	Explanation
CVGL	Cross-view geo-localization	MGTL	Mutual generative transformer learning
CAMask	Cascaded attention masking	CVI	Cross-view interaction
G2S	Ground-to-satellite	VIFE	View-independent feature extractor
S2G	Satellite-to-ground	SA	Spatial attention
SCE	Spatial context enhancement	MSFA	Multi-scale feature aggregation
GKST	Generative knowledge-supported transformer		

3.2.2. Feature Extractor

Formally, given an image pair $\mathbf{I}_G \in \mathbb{R}^{H_1 \times W_1 \times 3}$ and $\mathbf{I}_S \in \mathbb{R}^{H_2 \times W_2 \times 3}$, a multi-branch backbone (i.e., a Siamese-like VGG-based convolutional network with parameters $\Theta_{\mathtt{VIFE}}$) is used to extract features and generate spatial attention masks for each view simultaneously:

$$\mathbf{F}_G^{'} = f_{\mathtt{VIFE}}(\mathbf{I}_G; \Theta_{\mathtt{VIFE}}); \mathbf{F}_S^{'} = f_{\mathtt{VIFE}}(\mathbf{I}_S; \Theta_{\mathtt{VIFE}}) \tag{2}$$

where $\mathbf{F}_G^{'} \in \mathbb{R}^{c \times h \times w}$ and $\mathbf{F}_S^{'} \in \mathbb{R}^{c \times h \times w}$ are semantic representations with c channels and $h \times w$ spatial resolutions for ground-view and satellite-view, respectively.
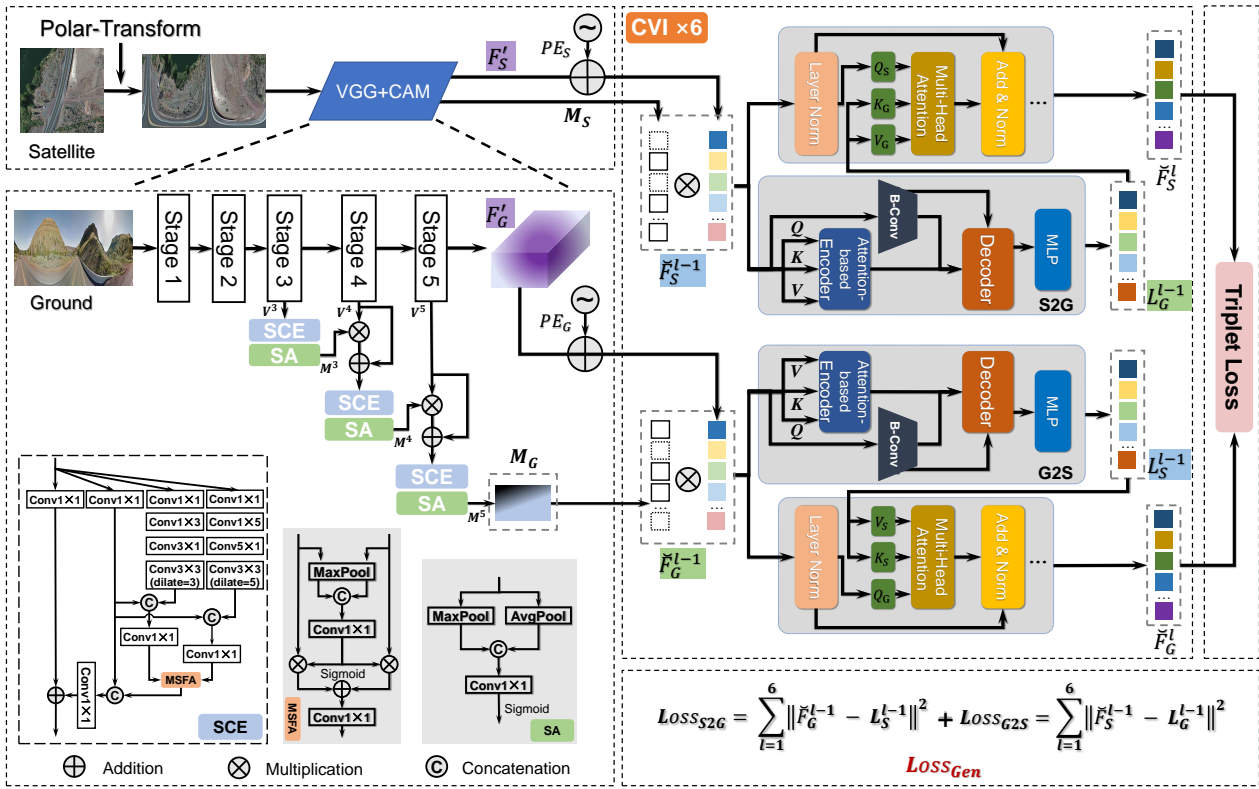
Figure 3. Overview of the proposed MGTL.

3.2.3. Cascaded Attention Masking

Viewpoint changes result in drastic appearance differences, which means much redundant information exists in $\mathbf{F}'_G$ and $\mathbf{F}'_S$ while matching. To encourage the network to focus on the co-visual regions, we designed a cascaded attention-masking (CAMask) algorithm and integrated it into the VGG16 backbone [67], seeking to learn the spatial attention masks that inhibit the non-co-visual areas adaptively. Figure 3 (left) illustrates the basic structure of the CAMask. Generally, the CAMask takes the side-output features $\{\mathbf{V}^i\}_{i=3}^5$ generated by the backbone as input and produces spatial attention masks $\mathbf{M}$ to enhance the inter-view co-visual information. Specifically, the fine-grained feature map captured by spatial context enhancement (SCE) is fed into two parallel pooling layers (i.e. maxpooling and avgpooling) along the channel dimension to generate two single-channel feature maps, respectively. Subsequently, these feature maps are concatenated along the channel dimension, and a convolutional layer is employed to adaptively generate masks with $h \times w$ resolutions. Spatial attention (SA) is illustrated in Figure 3. Note that, for the sake of brevity, $\mathbf{M}$ can refer to $\mathbf{M}_G$ or $\mathbf{M}_S$. The cascaded process can be formulated as follows:

$$\begin{aligned}
\mathbf{M}^3 &= \mathbf{SA}(\mathbf{SCE}(\mathbf{V}^3)), \\
\mathbf{M}^4 &= \mathbf{SA}(\mathbf{SCE}(\mathbf{V}^4 \otimes \mathbf{M}^3 + \mathbf{V}^4)), \\
\mathbf{M}^5 &= \mathbf{SA}(\mathbf{SCE}(\mathbf{V}^5 \otimes \mathbf{M}^4 + \mathbf{V}^5)),
\end{aligned} \tag{3}$$

where $\mathbf{M}^i$ represents the spatial attention mask of the i-th stage, $\mathbf{V}^i$ represents the feature map produced by the i-th stage in the backbone, and $\mathbf{M}^5 \in \mathbb{R}^{h \times w}$ is the final spatial attention mask $\mathbf{M}$. A better understanding of CAMask can be gained by focusing on its two components: spatial context enhancement (SCE) and spatial attention (SA).

Spatial Context Enhancement (SCE). To capture nuanced co-visual relationships, we meticulously devised a novel multi-branch convolutional module that effectively extracts fine-grained spatial representations from each view by utilizing diverse receptive fields. Fan [68] proposed a texture-enhanced module (TEM) consisting of multiple convolutional branches with different receptive fields. There is evidence that it facilitates the sensitive

capture of small spatial shifts. There are, however, certain limitations to coarse direct concatenation in TEM when the convolutional branches are independent. Motivated by this, we designed the SCE equipped with the multi-scale feature aggregation (MSFA) module to integrate branches with the guidance of spatial attention mechanism. As shown in Figure 3 (left), the SCE includes a shortcut branch and three parallel residual branches $\{b_i\}_{i=1}^3$ with different dilation rates $d \in \{1,3,5\}$, respectively. The shortcut branch utilizes a 1×1 convolutional layer to generate h_0 with channel size C. The branch b_1 only contains a 1×1 convolutional layer to halve the channel, while the remaining two branches $\{b_i\}_{i=2}^3$ adopt a 1×1 convolutional layer to reduce the channel and consist of three convolutional layers, i.e., a $1 \times (2i-1)$ convolutional layer, a $(2i-1) \times 1$ convolutional layer, and a 3×3 convolutional layer with dilation rate $(2i-1)$, to fully explore the spatial context information with rich receptive fields. Let $\{h_i\}_{i=1}^3$ represent the feature maps produced by the residual branches $\{b_i\}_{i=1}^3$, respectively. To fully explore the multi-scale information from the features $\{h_i\}_{i=1}^3$ generated by different convolutional layers, we carefully designed a **multi-scale feature aggregation (MSFA)** module by taking into account the specificities of spatial regions rather than concatenating them directly. Specifically, we concatenated the features $\{h_i\}_{i=2}^3$ with h_1 and fed the concatenated feature maps into a 1×1 convolutional layer to produce features $\{h_i'\}_{i=2}^3$ with unified channel C. MSFA takes $\{h_i'\}_{i=2}^3$ as input and produces attention-aware feature maps h_{msfa} that are then concatenated with h_1 followed by a 1×1 convolutional layer with a GeLU activation, then added up with h_0 to produce the final enhanced contextual feature representation.

Spatial Attention (SA). Inspired by [69], we learned the spatial attention masks according to the enhanced contextual representations adaptively. In detail, SA takes the enhanced feature produced by SCE and eliminates the channel dimension by adopting the maximum and average pooling layers. In order to generate the spatial attention masks $\mathbf{M}_G(\mathbf{M}_S) \in \mathbb{R}^{h \times w}$, we concatenate the compact features obtained from the pooling layers and then apply a 1×1 convolutional layer with sigmoid activation.

To alleviate the limitation of feature location on the receptive learning field, we re-encode the features $\mathbf{F}_G'$ and $\mathbf{F}_S'$ with position information and enrich the co-visual areas by multiplying the spatial attention masks $\mathbf{M}_G$ and $\mathbf{M}_S$ generated by the cascaded attention-masking (CAMask) algorithm:

$$\hat{\mathbf{F}}_G = (\mathbf{F}_G' + \text{PE}_G)\mathbf{M}_G; \hat{\mathbf{F}}_S = (\mathbf{F}_S' + \text{PE}_S)\mathbf{M}_S \tag{4}$$

where $\hat{\mathbf{F}}_G, \hat{\mathbf{F}}_S \in \mathbb{R}^{l \times c}$ are compact and position-aware feature representations and $l = h \times w$. Following [43], PE_G and PE_S are the positional encoding of feature maps $\mathbf{F}_G'$ and $\mathbf{F}_S'$, respectively.

3.3. Cross-View Synthesis

A key principle of our proposed mutual generative transformer learning (MGTL) is cross-view interaction (CVI), which is achieved by generating mutual simulated knowledge through cross-view generative modules f_{G2S} and f_{S2G} with the supervision of generative loss in Equation (10). We emphasize that the ground view cannot obtain the matched satellite view in advance during the evaluation/localization period, which makes it impossible to directly take one view as input and produce the features of another view in the training phase. Therefore, each generative module takes only the view feature from the self-branch as input to produce cross-view knowledge by using another view feature as supervision, which means that two sub-branches are completely decoupled while evaluating unlabeled image pairs. Generative modules are embedded in transformer layers, and the generative knowledge is utilized to calculate the *Key* and *Value* while performing the attention mechanism. Further, the generative module is trained with all the transformer parts to fully mine semantic consistency across views using the generative knowledge-supported transformer, which we called generative transformer learning. Figure 3 (right) illustrates the overview of the proposed cross-view interaction (CVI), whereby one view's information is taken

as input to generate knowledge that is aware of another view. The co-visual enhanced and position-aware representations $\hat{\mathbf{F}}_G$ and $\hat{\mathbf{F}}_S$ are further normalized as $\check{\mathbf{F}}_G = \text{LN}(\hat{\mathbf{F}}_G)$ and $\check{\mathbf{F}}_S = \text{LN}(\hat{\mathbf{F}}_S)$ to maintain representational capacity, respectively, where LN indicates the linear encoding operation following layer normalization. As shown in Figure 3 (right), the cross-view interaction module f_{CVI} is constructed by coupling two generative sub-modules f_{G2S} and f_{S2G} with an encoder–decoder structure as follows:

$$\mathbf{L}_S = f_{\text{G2S}}(\check{\mathbf{F}}_G), \mathbf{L}_G = f_{\text{S2G}}(\check{\mathbf{F}}_S). \tag{5}$$

Cross-View Generative Module f_{G2S} **and** f_{S2G}. Unet-like [70] architecture comprising an encoder and a decoder has been widely used in generative tasks recently. Existing sduties [25,42] demonstrate that the attention mechanism in a transformer is excellent at modeling global contextual information, and CNN excels at encoding local semantic information. With these properties in mind, we propose a novel generative module that owns Unet-like [70] architecture and combines multi-head self-attention and convolutional layers in parallel for mutual benefit. Taking f_{G2S} as an example, the hidden feature representation $\check{\mathbf{F}}_G$ is fed into the generative module to generate the simulated satellite-view feature representation $\mathbf{L}_G$, and the normalized satellite-view representation $\check{\mathbf{F}}_S$ is used for supervision, and vice versa. It is worth noting that both generative modules f_{G2S} and f_{S2G} own the same architecture but do not share weights due to the difference in input and generative content.

Encoder: Figure 4 illustrates the encoder–decoder architecture in detail. The encoder in the generative module is designed as a hybrid architecture that combines multi-head attention and convolutional layers. Taking f_{G2S} as an illustration, the feature $\check{\mathbf{F}}_G$ is encoded independently by the attention layers and the convolutional layers, resulting in producing the compact features $\dot{\mathbf{F}}_G^T$ and $\dot{\mathbf{F}}_G^C$, respectively, and these two features are concatenated along the channel to form the encoded feature $\dot{\mathbf{L}}_S$, which contains both global and local contextual information. **Decoder**: Following the acquisition of $\dot{\mathbf{L}}_S$, the decoding process begins with a two-layer multi-head attention operation followed by multi-layer perceptions to generate the simulated cross-view feature $\mathbf{L}_S$. The encoder and decoder are combined via skip connections to form a Unet-like [70] architecture, which enables aggregate features at different semantic levels.
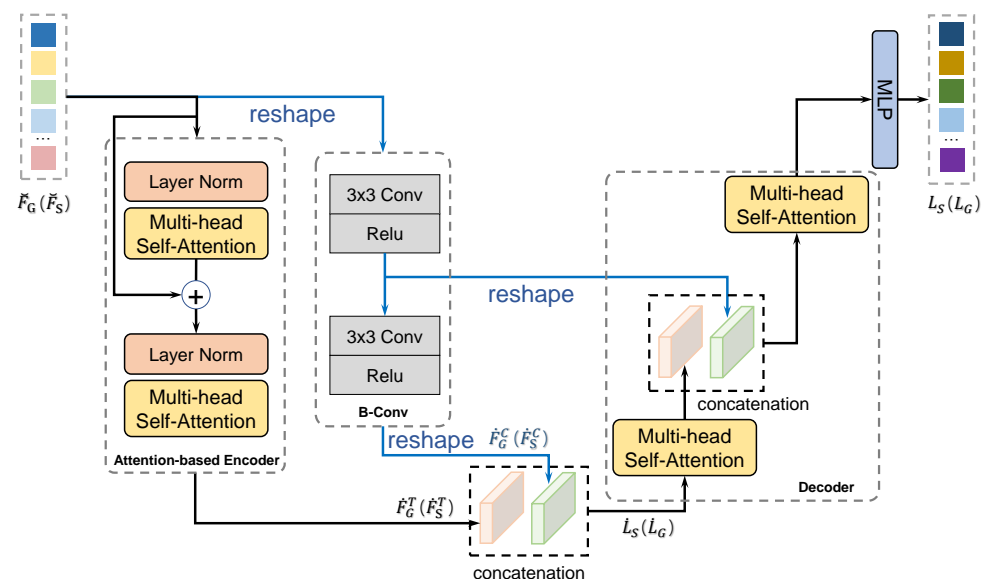
Figure 4. Details of the cross-view generative module. The generative module is designed as a Unet-like [70] architecture, taking advantage of transformer and CNN features to extract contextual information.

3.4. Generative Knowledge Supported Transformer (GKST) f_{GKST}

So far, we have acquired the inter-view representation $\check{\mathbf{F}}_G(\check{\mathbf{F}}_S)$ and the generative cross-view representation $\mathbf{L}_S(\mathbf{L}_G)$. To learn the final representation, $\mathbf{F}_G$ and $\mathbf{F}_S$, we designed a generative knowledge-supported transformer (GKST) to fully utilize all information. Formally, f_{GKST} takes $\check{\mathbf{F}}_G \in \mathbb{R}^{l \times c}(\check{\mathbf{F}}_S \in \mathbb{R}^{l \times c})$ and $\mathbf{L}_S \in \mathbb{R}^{l \times c}(\mathbf{L}_G \in \mathbb{R}^{l \times c})$ as inputs and produces the final high-order representations $\mathbf{F}_G(\mathbf{F}_S)$. Taking the ground view as an illustration, we feed the inter-view representation $\check{\mathbf{F}}_G$ and cross-view knowledge $\mathbf{L}_S$ into a multi-head cross-attention layer to learn the cross-view enhanced features. The cross-attention process is formulated as follows:

$$
\begin{aligned}
&Q_G^i = \check{\mathbf{F}}_G \mathbb{W}_Q^i, K_S^i = \mathbf{L}_S \mathbb{W}_K^i, V_S^i = \mathbf{L}_S \mathbb{W}_V^i, \\
&\texttt{Head}_i = \texttt{Attention}(Q_G^i, K_S^i, V_S^i), \\
&\texttt{MH}(Q, K, V) = \texttt{Concat}(\texttt{Head}_1, ..., \texttt{Head}_n)\mathbb{W},
\end{aligned}
\tag{6}
$$

where $\mathbb{W}_Q^i$, $\mathbb{W}_K^i$, $\mathbb{W}_V^i$, and $\mathbb{W}$ are learnable parameters. The updated representations $\hat{\mathbf{F}}_G$ can be achieved by two residual connections, which are formulated as follows:

$$
\begin{aligned}
\mathbf{F}_G^* &= \texttt{MH}(Q, K, V) + \check{\mathbf{F}}_G, \\
\hat{\mathbf{F}}_G &= \mathbf{F}_G^* + \texttt{LN}(\mathbf{F}_G^*).
\end{aligned}
\tag{7}
$$

We can easily obtain the final satellite-view feature maps $\hat{\mathbf{F}}_S$ in a similar way.

Recurrent Learning Process. To fully mine the benefits of the cross-view knowledge, we can further formulate the learning process recurrently as follows:

$$
\begin{cases}
\hat{\mathbf{F}}_G^l = f_{\text{GKST}}(\mathbf{L}_S^{l-1}, \check{\mathbf{F}}_G^{l-1}), \mathbf{L}_S^{l-1} = f_{\text{G2S}}(\check{\mathbf{F}}_G^{l-1}), \\
\hat{\mathbf{F}}_S^l = f_{\text{GKST}}(\mathbf{L}_G^{l-1}, \check{\mathbf{F}}_S^{l-1}), \mathbf{L}_G^{l-1} = f_{\text{S2G}}(\check{\mathbf{F}}_S^{l-1}),
\end{cases}
\tag{8}
$$

where $\check{\mathbf{F}}_G^{l-1} = \texttt{LN}(\hat{\mathbf{F}}_G^{l-1})$, $\check{\mathbf{F}}_S^{l-1} = \texttt{LN}(\hat{\mathbf{F}}_S^{l-1})$. Note that, at the beginning, ($l = 1$), $\hat{\mathbf{F}}_G^0$ and $\hat{\mathbf{F}}_S^0$ are produced by Equation (4), and the final representations $\mathbf{F}_G$ and $\mathbf{F}_S$ are produced by the last layer.

3.5. Loss Function

In order to make the final representations $\mathbf{F}_G$ and $\mathbf{F}_S$ more consistent between matching pairs but more discriminating among unmatching pairs, following [23], we employ a margin triplet loss $\text{Loss}_{\texttt{Triplet}}$ for final representation supervision:

$$
\text{Loss}_{\texttt{Triplet}} = \texttt{log}(1 + e^{\gamma(\text{d}_{pos} - \text{d}_{neg})}),
\tag{9}
$$

where γ indicates the function hyperparameter and d_{pos} and d_{neg} indicate the Euclidean distance between the positive and the negative pairs, respectively. To guarantee the quality of simulated cross-view knowledge, the cross-view knowledge generation module is supervised by mean squared errors (MSE) $\text{Loss}_{\texttt{Gen}}$:

$$
\text{Loss}_{\texttt{Gen}} = \sum_{l=1}^{L} (||\check{\mathbf{F}}_G^{l-1} - \mathbf{L}_S^{l-1}||^2 + ||\check{\mathbf{F}}_S^{l-1} - \mathbf{L}_G^{l-1}||^2),
\tag{10}
$$

where $\check{\mathbf{F}}_G^{l-1}(\check{\mathbf{F}}_S^{l-1})$ and $\mathbf{L}_S^{l-1}(\mathbf{L}_G^{l-1})$ denote inter-view representations and the generative cross-view knowledge at the l-th recurrent step, respectively. L is the total recurrent step. Finally, to learn the optimal parameters Θ for $\mathcal{F}_\Theta$, MGTL is jointly optimized through the overall learning Loss, which is computed as:

$$
\text{Loss} = \text{Loss}_{\texttt{Triplet}} + \lambda \text{Loss}_{\texttt{Gen}},
\tag{11}
$$

where λ is the balancing factor.

4. Experiments

4.1. Experimental Setting

Dataset: Following [14,22,23], we evaluated the performance of mutual generative transformer learning (MGTL) on two widely-used challenging benchmarks, *CVUSA* [30] and *CVACT* [31]. *CVUSA* was constructed by Workman et al. [9], containing 1.7 million training pairs collected from San Francisco. However, the relatively limited acquisition locations result in poor generalization capability of the extracted features when images from other positions are taken as input. To address this issue, the researchers reconstructed a new extensive *CVUSA* dataset [30], which contains 1.5 million geo-tagged pairs of ground-view and satellite-view images covering the continental United States, with resolutions of 1232×224 and 750×750, respectively. Ground-view images were collected using the Google Street View app and Flickr with different pre-processing methods. Specifically, the researchers randomly sampled images from the continental United States using the former but divided the entire area into a 100×100 grid and sampled up to 150 images in each cell while using the latter. Further, based on the original *CVUSA* [30], Zhai et al. [71] selected ground-view panoramas from *CVUSA* [30] and satellite-view images from Bing Maps at the same location as the matching pairs. In particular, the panoramas were wrapped to align with the satellite images using camera parameters. Finally, they released a subset of the CVUSA [30] containing 44,416 ground-satellite image pairs collected at the same location as well as 35,532 training pairs and 8884 evaluation pairs. This subset has become a widely-used benchmark because of its high resolution and simple format. To better investigate the possibility of matching geolocation in urban scenarios, Liu et al. [31] created a city-scale cross-view dataset *CVACT* [31] densely covering Canberra, Australia. Similar to *CVUSA*, ground-view panoramas were collected from the Google Street View app at zoom 2 with 1664×832 image resolution, while satellite-view images were collected from the Google Maps app at zoom 20 at the same location with 1200×1200 resolution. In order to fully evaluate the generalization of the CVGL methods, *CVACT* [31] released *CVACT*_test containing an extra 92,802 challenging pairs for testing only. Figure 5 displays several ground-satellite image pairs from *CVUSA* [30] and *CVACT* [31].

(a) CVUSA (b) CVACT

Figure 5. Example image pairs from *CVUSA* [30] and *CVACT* [32].

Evaluation Metric: Following existing works [14,23], recall accuracy at top K(r@K) was performed to evaluate the proposed MGTL. Staying in step with these existing methods, $K = 1, 5, 10, 1\%$ were selected.

Training Setting: During the training phase, MGTL adopted VGG16 [67] pre-trained on ImageNet [72] as the backbone. All training images were resized to 112×616 resolution augmented by random cropping, flipping, rotation, etc. We employed the Adam optimizer to optimize the whole network with the initial learning rate of 10^{-5}. We set the recurrent learning step of the generative knowledge supported transformer (GKST) to 6 and equipped 6 attention heads for each step. We set the batch size to 16 and trained the network for up to 150 epochs until complete convergence. The balancing factor λ in Equation (11) was carefully set to 0.05, and following [23], the regular item γ in Equation (9) was set to 10.0.

Reproducibility: We implemented the MGTL based on TensorFlow and trained the whole network on an NVIDIA GTX Titan X GPU with 12G CUDA memory.

4.2. Main Results

Baselines: Cross-view geo-localization (CVGL) has garnered significant research interest, resulting in several impressive works emerging in the field. To demonstrate the superiority of our proposed method, we selected 17 strong baselines and state-of-the-art methods in total, i.e., Workman et al. [9], Vo et al. [10], Zhai et al. [71], Cross-View Matching Network (CVM-Net) [11], Liu et al. [31], Regmi et al. [12], Spatial-Aware Feature Aggregation network (SAFA) [23], Cross-View Feature Transport technique (CVFT) [24], Dynamic Similarity Matching network (DSM) [73], Toker et al. [41], Layer-to-Layer Transformer (L2LTR) [14], Local Pattern Network (LPN) [26], Unit SAFA + Subtraction Attention Module (USAM) [74], LPN + USAM [74], pure transformer-based geo-localization (TransGeo) [13], Transformer-Guided Convolutional Neural Network (TransGCNN) [25], and LPN + Dynamic Weighted Decorrelation Regularization (DWDR) [27]. In particular, for omnidirectional comparison, we use their recommended settings for training. Our MGTL outperformed existing methods across most top $K(r@K)$ metrics on both benchmarks, showcasing the effectiveness of our proposed cascaded attention-masking (CAMask) algorithm and cross-view interaction (CVI) tactic. In this section, we provide a detailed introduction to our experiment setup and experiment results.

Performance on *CVUSA*: The test set of *CVUSA* [30] has 8884 challenging ground-satellite image pairs. The results with 17 SOTAs presented in Table 2 (left) show that our approach achieves state-of-the-art performance compared to all baselines in terms of almost all top $K(r@K)$ metrics on *CVUSA* [30]. Our approach achieves the best top 1(r@1) retrieval accuracy, as well as significant increases of 4.34% (90.16% → 94.50%) and 3.28% (91.22% → 94.50%) over SAFA+USAM [74] and LPN+USAM [74]. Notably, although the top 1% (r@1%) retrieval accuracy is almost 100%, MGTL still achieves 0.11% growth. Our approach outperforms L2LTR [14] by 0.45% (94.05% → 94.50%) in the top 1(r@1) retrieval accuracy while haaving less computation complexity and model capacity. TransGeo [13] utilizes a three-branch vision transformer [43] with a novel attention-based masking scheme. Our results outperform it by 0.42% (94.08% → 94.50%) in the top 1(r@1) retrieval accuracy. Nevertheless, methods other than MGTL ignore the semantic consistency revealed by cross-view interaction, making mutual generative learning the more convincing method. Figure 6 shows the partial hard image pair retrieval result from Toker et al. [41], L2LTR [14], SAFA [23], and CVFT [24]. The similarity between the ground truth and the selected unmatched satellite images heavily interferes with other models. In contrast, this illustrates that the co-visual enhanced features learned by CAMask and CVI own finer-grained understandings of scenarios and are highly discriminative.

Performance on *CVACT_val*: The evaluation set of *CVACT* [31] contains 8884 ground–satellite image pairs, consistent with *CVUSA* [30]. Table 2 (middle) presents the results with 13 STOAs on *CVACT*_val [31]. Our approach achieves the best performance across all top K(r@K) metrics (r@1, r@5, r@10, r@1%) on *CVACT*_val, i.e., 85.42%, 94.64%, 96.11%, and 98.51%, respectively. MGTL achieves significant improvements over LPN + USAM [74], SAFA + USAM [74], and LPN + DWDR [27], increasing the top 1(r@1) retrieval accuracy by 3.40% (82.02% → 85.42%), 3.02% (82.40% → 85.42%), and 1.69% (83.73% → 85.42%), respectively. In addition, our results are significantly better across all metrics compared to classical Siamese-like VGG-based convolutional methods, e.g., SAFA [23], CVFT [24], and DSM [73]. The experimental results mentioned above demonstrate the effectiveness of our CAMask and CVI introduced by MGTL. In comparison with traditional transformer-based methods TransGeo [13] and L2LTR [14], MGTL increased the top 1(r@1) retrieval accuracy significantly by 0.47% (84.95% → 85.42%) and 0.53% (84.89% → 85.42%), respectively. This strongly proves the superiority of our generative knowledge-supported transformer framework. Toker et al. [41] proposed a GAN-based method to synthesize realistic ground-view images from satellite images, which explores the benefits of generative learning for

cross-view matching. MGTL outperformed it by 2.14% in the top 1(r@1) retrieval accuracy, showing that our mutual generative learning strategy is more effective and has extreme generalizability in urban scenarios.
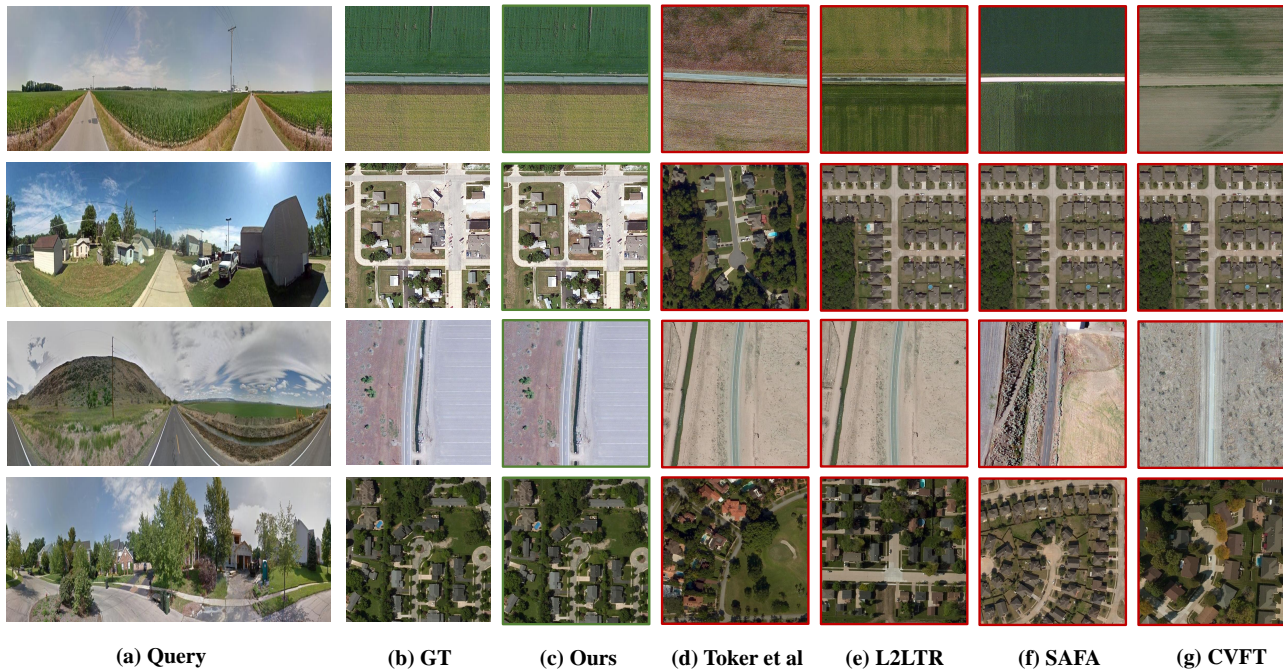

(a) Query   **(b) GT**   **(c) Ours**   **(d) Toker et al**   **(e) L2LTR**   **(f) SAFA**   **(g) CVFT**

Figure 6. Comparison results of some hard pairs.

Table 2. Quantitative results on the *CVUSA* [30] and *CVACT* [31] dataset.

Model	CVUSA				CVACT_val				CVACT_test			
	r@1	r@5	r@10	r@1%	r@1	r@5	r@10	r@1%	r@1	r@5	r@10	r@1%
2015 Workman et al. [9]	-	-	-	34.30	-	-	-	-	-	-	-	-
2016 Vo et al. [10]	-	-	-	63.70	-	-	-	-	-	-	-	-
2017 Zhai et al. [71]	-	-	-	43.20	-	-	-	-	-	-	-	-
2018 CVM-Net [11]	22.47	49.98	63.18	93.62	20.15	45.00	56.87	87.57	5.41	14.79	25.63	54.53
2019 Liu et al. [31]	40.79	66.82	76.36	96.12	46.96	68.28	75.48	92.04	19.9	34.82	41.23	63.79
2019 Regmi et al. [12]	48.75	-	81.27	95.98	-	-	-	-	-	-	-	-
2019 SAFA [23]	89.84	96.93	98.14	99.64	81.03	92.80	94.84	98.17	55.50	79.94	85.08	94.49
2020 CVFT [24]	61.43	84.69	90.49	99.02	61.05	81.33	86.52	95.93	34.39	58.83	66.78	95.99
2020 DSM [73]	91.96	97.50	98.54	99.67	82.49	92.44	93.99	97.32	35.55	60.17	67.95	86.71
2021 Toker et al. [41]	92.56	97.55	98.33	99.57	83.28	93.57	95.42	98.22	61.29	85.13	89.14	98.32
2021 L2LTR [14]	94.05	98.27	98.99	99.67	84.89	94.59	95.96	98.37	60.72	85.85	89.88	96.12
2021 LPN [26]	93.78	98.50	99.03	99.72	82.87	92.26	94.09	97.77	-	-	-	-
2022 SAFA+USAM [74]	90.16	-	-	99.67	82.40	-	-	98.00	56.16	-	-	95.22
2022 LPN+USAM [74]	91.22	-	-	99.67	82.02	-	-	98.18	37.71	-	-	87.04
2022 TransGeo [13]	94.08	98.36	99.04	99.77	84.95	94.14	95.78	98.37	-	-	-	-
2022 TransGCNN [25]	94.15	98.21	98.94	99.79	84.92	94.46	95.88	98.36	-	-	-	-
2022 LPN+DWDR [27]	94.33	**98.54**	99.09	**99.80**	83.73	92.78	94.53	97.78	-	-	-	-
Ours	**94.50**	98.41	**99.20**	99.78	**85.42**	**94.64**	**96.11**	**98.51**	**61.55**	**86.61**	**90.74**	**98.46** [1]

[1] Results are cited directly, and the best results are highlighted in bold.

Performance on *CVACT*_test: *CVACT*_test is massive and extremely challenging, consisting of 92,802 ground–satellite image pairs in urban scenarios for testing only. For the challenging *CVACT*_test, we compared our approach with 9 SOTAs. As shown in Table 2 (Right), our MGTL sets new retrieval accuracy records across all metrics compared to existing SOTAs. MGTL increases the top 1(r@1) retrieval accuracy significantly by

0.83% (60.72% → 61.55%) and 5.39% (56.16% → 61.55%) compared to L2LTR [14] and SAFA+USAM [74], respectively. Furthermore, our results not only outperform others in top 1(r@1) retrieval accuracy, but also gain a remarkable increase of 1.48% (85.13% → 86.61%) in the top 5(r@5) retrieval accuracy over Toker et al. [41] and 2.34% (96.12% → 98.46%) in the top 1%(r@1%) recall accuracy over L2LTR [14]. These superior experiment results showcase that MGTL is capable of capturing high-order understandings of cross-view scenarios essential for CVGL in unfamiliar environments in the absence of prior knowledge.

4.3. Ablation Study

As the mutual generative transformer learning (MGTL) incorporates the cascaded attention-masking (CAMask) algorithm and cross-view interaction (CVI) tactic into the cross-view geo-localization (CVGL) task, we conduct substantial ablation studies to carefully scrutinize how each component affects the learning ability of the model.

Effectiveness of CAMask: To qualitatively study the effectiveness of our proposed CAMask algorithm, we inspect the performance of the VGG16 backbone [67] with fully-connected layers removed. As shown in Table 3, all metrics degrade significantly when removing CAMask. The top 1(r@1) retrieval accuracy suffers a drastic decrease of 10.18%, from 90.12% to 79.94%, supporting the notion that co-visual information explicitly learned by CAMask is extremely critical for CVGL. In addition, we removed the CAMask from the fully equipped model. Observing the last two lines in Table 3, the top 1(r@1) retrieval accuracy still suffers a heavy decrease of 4.06%, from 94.50% to 90.44%, suggesting once again that the above notion is strongly supported. To explore the necessity of SCE and SA, Table 4 displays the comparison results while removing each of them, respectively. Replacing the SCE with fully convolutional blocks, the top 1(r@1) retrieval accuracy decreased by 2.21% (94.50% → 92.29%) and 3.05% (85.42% → 82.37%) on *CVUSA* and *CVACT_val*, respectively. Similarly, when we replaced the SA with global average pooling (GAP), the top 1(r@1) retrieval accuracy degraded by 0.88% (94.50% → 93.62%) and 1.09% (85.42% → 84.33%) on both datasets. MGTL suffers a decrease in precision for varying content, suggesting co-visual enhanced feature representations learned by CAMask lead to more reliable results. To show the superiority of CAMask qualitatively, we meticulously visualized the cascaded attention masks in Figure 7 to support our claim. The first row indicates the generative attention masks as well as corresponding attention scores. To showcase the co-visual regions intuitively, we binarize the attention masks, as shown in the second row. Subsequently, the original images are cropped with the guidance of binary masks, as shown in the third row. Observing the third row, only the co-visual regions (e.g., road, building) remain, and redundant non-co-visual regions (e.g., 'sky' in ground imagery but absent in satellite imagery) useless for matching were masked. CAMask eradicates these disturbances in a simple but effective manner. Finally, to showcase the correctness of the co-visual relationships, the same regions captured across views are marked with rectangles of the same color, as shown in the fourth row.

Table 3. **Ablation study** of the proposed cascaded attention-masking (CAMask) algorithm.

Candidate			Complexity		CVUSA				CVACT_val			
VGG16	CAMask	CVI	GFLOPs ↓	Param. ↓	r@1	r@5	r@10	r@1%	r@1	r@5	r@10	r@1%
✔			28.22	29.42 M	79.94	93.66	96.25	99.31	70.67	87.73	91.13	95.78
✔	✔		59.02	91.18 M	85.15	95.13	96.89	99.43	76.32	89.64	92.26	96.21
✔		✔	28.90	137.21 M	90.44	96.83	97.41	99.45	81.25	92.12	94.38	97.69
✔	✔	✔	59.71	171.97 M	**94.50**	**98.41**	**99.20**	**99.78**	**85.42**	**94.64**	**96.11**	**98.51** [1]

[1] The best results are highlighted in bold. ↓ means lower is better.

Table 4. Ablation study of the proposed spatial attention (SA) and spatial context enhancement (SCE) in cascaded attention masking (CAMask).

VGG16 + CVI	Candidate		Complexity		CVUSA				CVACT_val			
	w/SA	w/SCE	GFLOPs ↓	Param. ↓	r@1	r@5	r@10	r@1%	r@1	r@5	r@10	r@1%
✔			28.90	137.21 M	90.44	96.83	97.41	99.45	81.25	92.12	94.38	97.69
✔	✔		28.90	137.21 M	92.29	97.65	98.67	99.72	82.37	93.35	95.17	98.23
✔		✔	59.71	171.97 M	93.62	98.41	99.07	99.73	84.33	94.31	95.67	98.43
✔	✔	✔	59.71	171.97 M	**94.50**	**98.41**	**99.20**	**99.78**	**85.42**	**94.64**	**96.11**	**98.51** [1]

[1] 'w/' means the proposed MGTL is equipped with SA or SCE, respectively. The best results are highlighted in bold. ↓ means lower is better.
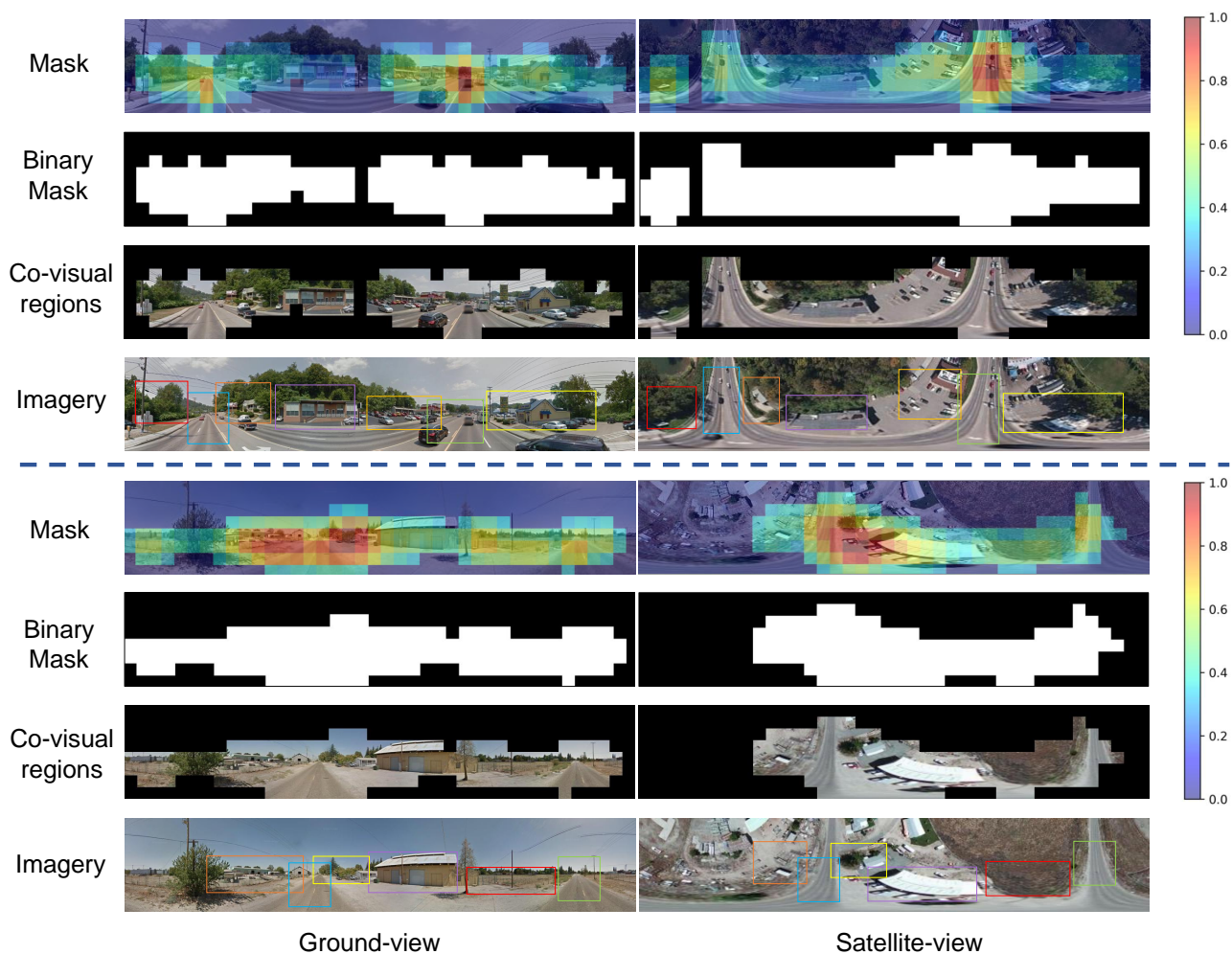
Figure 7. Visualization results of cascaded attention masks.

Effectiveness of CVI: We introduce the CVI tactic to implicitly explore co-visual information and empirically investigate the superiority of CVI in MGTL. Table 5 shows that when incorporating transformer [42] blocks, the top 1(r@1) retrieval accuracy increases by 6.27% (85.15% → 91.42%). However, this is still a sub-optimal performance compared to existing SOTAs. The designation 'w/o' in Table 5 refers to the pure transformer without CVI; all metrics suffer drastic degradations, and the top 1(r@1) retrieval accuracy decreases by 3.08% (94.50% → 91.42%), showcasing that CVI further boosts the pure transformer learning capability and enhances the similarity of feature representations between matching pairs.

Table 5. **Ablation study** of cross-view interaction (CVI).

VGG16 + CAMask	Candidate		Complexity		CVUSA				CVACT_val			
	w/o	w/CVI	GFLOPs ↓	Param. ↓	r@1	r@5	r@10	r@1%	r@1	r@5	r@10	r@1%
✔			59.02	91.98 M	85.15	95.13	96.89	99.43	76.32	89.64	92.26	96.21
✔	✔		59.32	113.48 M	91.42	96.21	98.04	99.62	81.99	93.16	95.04	98.23
✔		✔	59.71	171.97 M	**94.50**	**98.41**	**99.20**	**99.78**	**85.42**	**94.64**	**96.11**	**98.51** [1]

[1] 'w/' and 'w/o' means the transformer learning is or is not equipped with CVI, respectively. The best results are highlighted in bold. ↓ means lower is better.

Composition of Generative Module: To determine the most effective interaction mode, we empirically explored the variational autoencoder (VAE) [75], CNN-based Unet [70], and pure transformer [42] block. The results shown in Table 6 suggest the superiority of our hybrid generative module. Specifically, VAE [75] is considered to be one of the classical generative models, so we utilized 2 two-layer fully-connected networks as encoder and decoder, respectively. Following [70], we exploited 2 two-layer convolutional blocks as both encoder and decoder to form a simplified Unet-like [70] architecture. However, limited by the locality assumptions, there is significant deterioration in precision. Similarly, we reconstructed a simplified transformer-based generative module referring to [47], whose encoder and decoder both consisted of two transformer layers. In this case, the performance across all metrics was worse than ours, but the complexity is higher. These results demonstrate that our generative module is suitable for plugging into a transformer to generate simulated cross-view knowledge.

Table 6. **Detailed ablation study** of the composition of the generative module.

Method	Complexity		CVUSA				CVACT_val			
	GFLOPs ↓	Param. ↓	r@1	r@5	r@10	r@1%	r@1	r@5	r@10	r@1%
VAE [75]	59.89	141.84M	94.11	98.27	99.03	99.71	85.32	94.42	96.04	98.41
Unet [70]	59.53	134.74M	92.04	97.91	98.80	99.67	82.31	93.08	95.09	98.30
Transformer [42]	61.20	276.49M	94.37	98.30	99.08	99.74	85.35	94.45	96.03	98.44
Ours	59.71	171.97M	**94.50**	**98.41**	**99.20**	**99.78**	**85.42**	**94.64**	**96.11**	**98.51** [1]

[1] The best results are highlighted in bold. ↓ means lower is better.

Study of Recurrent Learning Steps: To determine the best recurrent learning step that balances quality and complexity, we report the results trained with different recurrent steps in Table 7. This demonstrates that increasing the recurrent learning step leads to improved performance, which proves that recurrent learning can fully mine the representational ability of generative cross-view knowledge. As the learning step increases from 3 to 6, the top 1(r@1) retrieval accuracy improves significantly. However, the gains are negligible and even degrade as the recurrent step rises to 9. We note that the quantity and quality of the new generative knowledge becomes more difficult as the recurrent step rises. Therefore, the recurrent learning step is set to 6 to achieve a trade-off between accuracy and time cost.

Table 7. **Detailed ablation study** of different parameter settings.

CVI	CVUSA				CVACT_val			
	r@1	r@5	r@10	r@1%	r@1	r@5	r@10	r@1%
L = 1	87.07	96.48	97.72	99.65	77.92	90.85	93.27	97.20
L = 3	89.67	97.15	98.26	99.70	80.40	91.63	94.18	98.17
L = 6	**94.50**	**98.41**	**99.20**	**99.78**	**85.42**	**94.64**	96.11	**98.51**
L = 9	94.25	98.39	99.18	99.76	85.20	94.61	**96.12**	98.49 [1]

[1] The best results are highlighted in bold.

4.4. Supplementary Experiment

To further explore the rationality of each module, i.e., multi-scale feature aggregation (MSFA), spatial context enhancement (SCE), and spatial attention (SA) in the cascaded attention-masking (CAMask) algorithm, we conduct intensive experiments with different settings on both *CVUSA* [30] and *CVACT_val* [31]. All results are reported in Table 8.

Exploration of MSFA: We redesign a parallel multi-branch convolutional module named SCE. Unlike existing works, we introduce a novel feature aggregation tactic named MSFA to further integrate branches with different scales. To prove the necessity of MSFA, we replace the MSFA with direct addition operations and convolutional layers, respectively. We observe that all the metrics decreased and that the top 1(r@1) retrieval accuracy decreased drastically by more than 1%. To further illustrate the advancements of our proposed MSFA, we select five typical feature aggregation tactics, including squeeze-and-excitation networks (SENet) [76], convolutional block attention module (Cbam) [69], self-calibrated convolution (SCNet) [77], Non_local [78], and selective kernel convolution (SKC) [79], and then plug them into SCE. As shown in Table 8 (top), SCE with MSFA achieves the best retrieval accuracy, with a parameter increase of less than 10 M.

Exploration of SCE: To illustrate the superiority of our proposed SCE, we select two typical multi-branch convolutional modules, i.e., Texture-Enhanced Module (TEM) [68] and Receptive Field Block (RFB) [80]. TEM [68] aims to capture fine-grained texture and context features, and it was initially employed in the concealed object detection (COD) task. Inspired by the human visual system, RFB [80] introduced a multi-branch dilated convolution to enhance the feature extraction ability of the network. Building on this purpose, we introduce SCE with MSFA. As shown in Table 8 (middle), SCE achieves the optimal performance with fewer parameters, suggesting the SCE equipped with an attention-based feature aggregation tactic is more suitable for the CVGL task.

Table 8. **Detailed ablation study** of the rationality of MSFA and MSCM.

Module	Complexity		CVUSA				CVACT_val			
	GFLOPs ↓	Param. ↓	r@1	r@5	r@10	r@1%	r@1	r@5	r@10	r@1%
Why MSFA in SCE?—Comparison with other aggregation methods.										
add	54.62	161.30 M	93.11	98.22	98.93	99.75	83.55	93.66	95.56	98.26
concat + conv	55.75	163.71 M	92.62	97.96	98.85	99.72	83.07	93.84	95.57	98.21
SENet [76]	54.63	161.95 M	93.48	98.31	98.99	99.74	85.01	94.40	96.02	98.48
Cbam [69]	54.63	161.96 M	93.61	98.39	99.02	99.73	84.89	94.27	96.03	98.43
SCNet [77]	56.68	166.12 M	93.82	98.38	99.08	99.74	84.67	94.28	95.97	98.35
Non_Local [78]	65.36	163.72 M	92.84	98.01	98.86	99.67	83.48	93.99	95.68	98.29
SKC [79]	73.84	201.91 M	93.92	98.36	99.03	99.78	84.95	94.36	95.81	98.44
MSFA	59.71	171.97 M	**94.50**	**98.41**	**99.20**	**99.78**	**85.42**	**94.64**	**96.03**	**98.51**
Why SCE in CAMask?—Comparison with other parallel multi-branch convolutional modules.										
Conv	76.38	208.51 M	92.97	98.16	98.87	99.72	83.51	94.02	95.78	98.37
TEM [68]	67.05	187.31 M	94.21	98.37	99.05	99.74	85.10	94.56	95.99	98.43
RFB [80]	67.61	188.47 M	94.13	98.33	99.07	99.72	85.08	94.48	96.01	98.40
SCE	59.71	171.97 M	**94.50**	**98.41**	**99.20**	**99.78**	**85.42**	**94.64**	**96.03**	**98.51**
Why SA in CAMask?—Comparison with other attention mask generation methods.										
GAP	59.71	171.97 M	93.62	98.41	99.07	99.73	84.33	94.31	95.67	98.43
GMP	59.71	171.97 M	93.43	98.33	99.05	99.75	84.20	94.21	95.88	98.45
SA	59.71	171.97 M	**94.50**	**98.41**	**99.20**	**99.78**	**85.42**	**94.64**	**96.03**	**98.51** [1]

[1] The best results are highlighted in bold. ↓ means lower is better.

Exploration of SA: Spatial attention is thought to adaptively learn discriminative regions in the feature map to generate the spatial masks. We employ SA to connect cascaded structures and used spatial masks generated from cross-level semantic information to

compensate for the loss of spatial information due to reduced spatial resolutions via multiplying with high-level semantic features. To study the effectiveness of SA, we replace SA with global average/max pooling (GAP/GMP) layers and observe an overall decrease across all metrics in which the top 1(r@1) retrieval accuracy suffers drastic decreases by 0.88% (94.50% → 93.62%) and 1.07% (94.50% → 93.43%), showcasing the necessity and effectiveness of the SA mechanism in CAMask.

5. Discussion

As described in Section 4, mutual generative transformer learning (MGTL) outperforms recent outstanding cross-view geo-localization (CVGL) works significantly across almost all metrics on widely-used benchmarks *CVUSA* [30] and *CVACT* [31], owing to our cascaded attention-masking (CAMask) algorithm and cross-view interaction (CVI) tactic. CAMask is integrated into feature extractor VGG16 [67] to encourage co-visual regions for reasoning during generative transformer learning, which eradicates the interference of viewpoint-sensitive regions. CVI is implemented by a cross-view generative module and by generative knowledge-supported transformer learning. Cross-view mutual generative learning aims to simulate feature representations across views, subsequently exploiting generative knowledge to mine the semantic consistency through the attention mechanism in recurrent transformer learning. Our findings perform excellently in cross-view image matching essential for CVGL. In addition, our MGTL enhances the generalizability of CVGL, driving vision-based geo-localization solutions applicable in autonomous driving fields without GPS support.

6. Future Work

By exploiting the inter-view semantic consistency, mutual learning can alleviate ambiguity in cross-view matching. The study of view matching in UAV localization has also been conducted in a similar area for the purpose of completing UAV geographic localization. A benchmark called University-1652 [81] aims to establish correspondence between a UAV view and a satellite view. In our future work, we will explore how mutual learning techniques can play a role in this similar field, including: (1) The slight difference in view perspective between the satellite view and UAV view makes the cross-view semantic consistency easier to obtain, allowing mutual learning to go further in enhancing semantics; and (2) Shared parameter learning, which can make the network more efficient, should be explored in the context of mutual learning.

7. Conclusions

This paper proposed a novel mutual generative transformer learning network, denoted as MGTL, for addressing the cross-view geo-localization problem. Existing methods commonly rely on a CNN-based Siamese-like backbone to extract high-order feature representations and treat each region equally. Viewpoint-sensitive regions with drastic appearance differences, however, hinder image matching significantly. Using a cascaded attention-masking algorithm, we introduced a spatial context enhancement module and a spatial attention module in the VGG16 to capture co-visual information. In terms of semantic consistency learning, it is rarely examined in recent works, but incorporating consistency constraints by cross-view interaction during the recurrent learning process will benefit similarity computing. To facilitate high-order information mining within each view, we constructed cross-view generative modules and injected their generative cross-view knowledge into a transformer-based framework. Extensive qualitative and quantitative experiments demonstrated that mutual generative transformer learning significantly alleviated the impact of spatial information mismatch caused by drastic viewpoint changes. By examining cross-view interactions, we highlighted the potential of this perspective to advance automobile geolocation identification research in GPS-denied conditions.

References

1. Saurer, O.; Baatz, G.; Köser, K.; Pollefeys, M.; Ladicky, L.U. Image based geo-localization in the alps. *Int. J. Comput. Vis.* **2016**, *116*, 213–225. [CrossRef]
2. Senlet, T.; Elgammal, A. Satellite image-based precise robot localization on sidewalks. In Proceedings of the IEEE International Conference on Robotics and Automation, St Paul, MN, USA, 14–19 May 2012; pp. 2647–2653.
3. Xiao, Y.; Codevilla, F.; Gurram, A.; Urfalioglu, O.; López, A.M. Multimodal end-to-end autonomous driving. *IEEE Trans. Intell. Transp. Syst.* **2020**, *23*, 537–547. [CrossRef]
4. Wang, S.; Zhang, Y.; Li, H. Satellite image based cross-view localization for autonomous vehicle. *arXiv* **2022**, arXiv:2207.13506.
5. Thoma, J.; Paudel, D.P.; Chhatkuli, A.; Probst, T.; Gool, L.V. Mapping, localization and path planning for image-based navigation using visual features and map. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 7383–7391.
6. Roy, N.; Debarshi, S. Uav-based person re-identification and dynamic image routing using wireless mesh networking. In Proceedings of the 2020 7th International Conference on Signal Processing and Integrated Networks (SPIN) IEEE, Noida, Delhi-NCR, India, 27–28 February 2020; pp. 914–917.
7. Hu, S.; Lee, G.H. Image-based geo-localization using satellite imagery. *IJCV* **2020**, *128*, 1205–1219. [CrossRef]
8. Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; Sivic, J. NetVLAD: CNN architecture for weakly supervised place recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5297–5307.
9. Workman, S.; Jacobs, N. On the location dependence of convolutional neural network features. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 8–10 June 2015; pp. 70–78.
10. Vo, N.N.; Hays, J. Localizing and orienting street views using overhead imagery. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 494–509.
11. Hu, S.; Feng, M.; Nguyen, R.M.; Lee, G.H. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7258–7267.
12. Regmi, K.; Shah, M. Bridging the domain gap for ground-to-aerial image matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 470–479.
13. Zhu, S.; Shah, M.; Chen, C. 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, New Orleans, LA, USA, 19–23 June 2022; pp. 1162–1171.
14. Yang, H.; Lu, X.; Zhu, Y. Cross-view Geo-localization with Layer-to-Layer Transformer. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 29009–29020.
15. Chen, Z.; Lam, O.; Jacobson, A.; Milford, M. Convolutional neural network-based place recognition. *arXiv* **2014**, arXiv:1411.1509.
16. Xin, Z.; Cai, Y.; Lu, T.; Xing, X.; Cai, S.; Zhang, J.; Yang, Y.; Wang, Y. Localizing Discriminative Visual Landmarks for Place Recognition. In Proceedings of the IEEE International Conference on Robotics and Automation, Montreal, QC, Canada, 20–24 May 2019; pp. 5979–5985.
17. Khaliq, A.; Milford, M.; Garg, S. MultiRes-NetVLAD: Augmenting Place Recognition Training with Low-Resolution Imagery. *IEEE Robot. Autom. Lett.* **2022**, *7*, 3882–3889. [CrossRef]
18. Yu, J.; Zhu, C.; Zhang, J.; Huang, Q.; Tao, D. Spatial pyramid-enhanced NetVLAD with weighted triplet loss for place recognition. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *31*, 661–674. [CrossRef]

19. Latif, Y.; Garg, R.; Milford, M.; Reid, I. Addressing challenging place recognition tasks using generative adversarial networks. In Proceedings of the International Conference on Robotics and Automation, Brisbane, Australia, 21–26 May 2018; pp. 2349–2355.

20. Castaldo, F.; Zamir, A.; Angst, R.; Palmieri, F.; Savarese, S. Semantic cross-view matching. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Santiago, Chile, 7–13 December 2015; pp. 9–17.

21. Mousavian, A.; Kosecka, J. Semantic Image Based Geolocation Given a Map. *arXiv* **2016**, arXiv:1609.00278.

22. Zhu, S.; Yang, T.; Chen, C. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Online, 19–25 June 2021; pp. 3640–3649.

23. Shi, Y.; Liu, L.; Yu, X.; Li, H. Spatial-aware feature aggregation for image based cross-view geo-localization. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Volume 32.

24. 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, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11990–11997.

25. Wang, T.; Fan, S.; Liu, D.; Sun, C. Transformer-Guided Convolutional Neural Network for Cross-View Geolocalization. *arXiv* **2022**, arXiv:2204.09967.

26. 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]

27. Wang, T.; Zheng, Z.; Zhu, Z.; Gao, Y.; Yang, Y.; Yan, C. Learning Cross-view Geo-localization Embeddings via Dynamic Weighted Decorrelation Regularization. *arXiv* **2022**, arXiv:2211.05296.

28. Zhu, Y.; Yang, H.; Lu, Y.; Huang, Q. Simple, Effective and General: A New Backbone for Cross-view Image Geo-localization. *arXiv* **2023**, arXiv:2302.01572.

29. Zhang, X.; Li, X.; Sultani, W.; Zhou, Y.; Wshah, S. Cross-view Geo-localization via Learning Disentangled Geometric Layout Correspondence. *arXiv* **2022**, arXiv:2212.04074.

30. Workman, S.; Souvenir, R.; Jacobs, N. Wide-area image geolocalization with aerial reference imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015; pp. 3961–3969.

31. Liu, L.; Li, H. Lending orientation to neural networks for cross-view geo-localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5624–5633.

32. Zhu, Y.; Sun, B.; Lu, X.; Jia, S. Geographic Semantic Network for Cross-View Image Geo-Localization. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–15. [CrossRef]

33. Zhu, B.; Yang, C.; Dai, J.; Fan, J.; Ye, Y. R2FD2: Fast and Robust Matching of Multimodal Remote Sensing Image via Repeatable Feature Detector and Rotation-invariant Feature Descriptor. *IEEE Trans. Geosci. Remote Sens.* **2023**. [CrossRef]

34. Regmi, K.; Borji, A. Cross-view image synthesis using conditional gans. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3501–3510.

35. Lu, X.; Li, Z.; Cui, Z.; Oswald, M.R.; Pollefeys, M.; Qin, R. Geometry-aware satellite-to-ground image synthesis for urban areas. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 859–867.

36. Ding, H.; Wu, S.; Tang, H.; Wu, F.; Gao, G.; Jing, X.Y. Cross-view image synthesis with deformable convolution and attention mechanism. In Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Nanjing, China, 16–18 October 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 386–397.

37. Lin, T.Y.; Cui, Y.; Belongie, S.; Hays, J. Learning deep representations for ground-to-aerial geolocalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015; pp. 5007–5015.

38. Sun, B.; Chen, C.; Zhu, Y.; Jiang, J. GeoCapsNet: Aerial to Ground view Image Geo-localization using Capsule Network. *arXiv* **2019**, arXiv:1904.06281.

39. Cai, S.; Guo, Y.; Khan, S.; Hu, J.; Wen, G. Ground-to-Aerial Image Geo-Localization With a Hard Exemplar Reweighting Triplet Loss. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8391–8400.

40. Ren, B.; Tang, H.; Sebe, N. Cascaded cross mlp-mixer gans for cross-view image translation. *arXiv* **2021**, arXiv:2110.10183.

41. Toker, A.; Zhou, Q.; Maximov, M.; Leal-Taixé, L. Coming down to earth: Satellite-to-street view synthesis for geo-localization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Online, 11–17 October 2021; pp. 6488–6497.

42. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.

43. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.

44. Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; Gao, W. Pre-trained image processing transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Online, 19–25 June 2021; pp. 12299–12310.

45. Bhojanapalli, S.; Chakrabarti, A.; Glasner, D.; Li, D.; Unterthiner, T.; Veit, A. Understanding robustness of transformers for image classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Online, 11–17 October 2021; pp. 10231–10241.

46. Lanchantin, J.; Wang, T.; Ordonez, V.; Qi, Y. General multi-label image classification with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Online, 19–25 June 2021; pp. 16478–16488.

47. Strudel, R.; Pinel, R.G.; Laptev, I.; Schmid, C. Segmenter: Transformer for Semantic Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Online, 11–17 October 2021; pp. 7262–7272.

48. Jin, Y.; Han, D.; Ko, H. Trseg: Transformer for semantic segmentation. *Pattern Recognit. Lett.* **2021**, *148*, 29–35. [CrossRef]

49. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.; et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021; pp. 6881–6890.

50. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Online, August 23-28, 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.

51. Misra, I.; Girdhar, R.; Joulin, A. An end-to-end transformer model for 3d object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Online, 11–17 October 2021; pp. 2906–2917.

52. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.

53. Liang, Z.; Wang, Y.; Wang, L.; Yang, J.; Zhou, S. Light field image super-resolution with transformers. *IEEE Signal Process. Lett.* **2022**, *29*, 563–567. [CrossRef]

54. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.H. Restormer: Efficient transformer for high-resolution image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–23 June 2022; pp. 5728–5739.

55. Li, Z.; Liu, X.; Drenkow, N.; Ding, A.; Creighton, F.X.; Taylor, R.H.; Unberath, M. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Online, 11–17 October 2021; pp. 6197–6206.

56. Ding, Y.; Yuan, W.; Zhu, Q.; Zhang, H.; Liu, X.; Wang, Y.; Liu, X. Transmvsnet: Global context-aware multi-view stereo network with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–23 June 2022; pp. 8585–8594.

57. He, X.; Chen, Y.; Lin, Z. Spatial-spectral transformer for hyperspectral image classification. *Remote Sens.* **2021**, *13*, 498. [CrossRef]

58. Qing, Y.; Liu, W.; Feng, L.; Gao, W. Improved transformer net for hyperspectral image classification. *Remote Sens.* **2021**, *13*, 2216. [CrossRef]

59. Sun, L.; Zhao, G.; Zheng, Y.; Wu, Z. Spectral-spatial feature tokenization transformer for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]

60. Zhou, H.; Tian, C.; Zhang, Z.; Huo, Q.; Xie, Y.; Li, Z. Multispectral fusion transformer network for RGB-thermal urban scene semantic segmentation. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]

61. Chu, X.; Tian, Z.; Zhang, B.; Wang, X.; Wei, X.; Xia, H.; Shen, C. Conditional positional encodings for vision transformers. *arXiv* **2021**, arXiv:2102.10882.

62. Li, Y.; Zhang, K.; Cao, J.; Timofte, R.; Van Gool, L. Localvit: Bringing locality to vision transformers. *arXiv* **2021**, arXiv:2104.05707.

63. Chen, C.F.R.; Fan, Q.; Panda, R. Crossvit: Cross-attention multi-scale vision transformer for image classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Online, 11–17 October 2021; pp. 357–366.

64. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Online, 11–17 October 2021; pp. 10012–10022.

65. Yang, F.; Zhai, Q.; Li, X.; Huang, R.; Luo, A.; Cheng, H.; Fan, D.P. Uncertainty-guided transformer reasoning for camouflaged object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Online, 11–17 October 2021; pp. 4146–4155.

66. Wang, W.; Yao, L.; Chen, L.; Cai, D.; He, X.; Liu, W. CrossFormer: A Versatile Vision Transformer Based on Cross-scale Attention. *arXiv* **2021**, arXiv:2108.00154.

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

68. Fan, D.P.; Ji, G.P.; Cheng, M.M.; Shao, L. Concealed object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 6024–6042. [CrossRef] [PubMed]

69. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the IEEE European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 3–19.

70. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.

71. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 867–875.

72. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.

73. 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, Seattle, WA, USA, 13–19 June 2020; pp. 4064–4072.

74. 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]
75. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
76. Jie, H.; Li, S.; Gang, S. Squeeze-and-Excitation Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
77. Liu, J.J.; Hou, Q.; Cheng, M.M.; Wang, C.; Feng, J. Improving Convolutional Networks With Self-Calibrated Convolutions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 10096–10105.
78. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803.
79. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective Kernel Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 510–519.
80. Liu, S.; Huang, D.; Wang, Y. Receptive field block net for accurate and fast object detection. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 385–400.
81. 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, Seattle, WA, USA, 12–16 October 2020; pp. 1395–1403.