Homography Estimation With Adaptive Query Transformer and Gated Interaction Module

Zhongyang Li, Faming Fang, Tingting Wang, and Guixu Zhang

***Abstract*— Homography estimation is essential for aligning images captured from different viewpoints by accurately modeling the geometric relationship between them. In homography estimation, global information plays a critical role. To establish global correspondences, cross-attention has been widely used in recent studies. However, vanilla cross-attention mechanisms treat queries in redundant and low-texture areas the same as those in richly textured areas, leading to the accumulation and propagation of erroneous information. We define this phenomenon, where the model excessively attends to queries in redundant and low-texture areas, as *query over-focusing*. To alleviate query over-focusing and achieve fine-grained homography estimation, we propose a novel homography estimation network, termed AGNet, which integrates an Adaptive Query Transformer (AQFormer) and a Gated Interaction Module (GIM). The AQFormer is designed to dynamically adjust attention by applying a mask to queries, allowing the model to adaptively emphasize feature-rich regions while suppressing redundant or weakly textured areas. Meanwhile, the GIM selectively captures local information by adjusting convolutional kernels based on input, enhancing the extraction of shared features between image pairs. Extensive experiments on various datasets demonstrate that AGNet significantly improves accuracy in homography estimation, particularly in challenging scenarios with low overlap and large viewpoint variations.**

***Index Terms*— Deep learning, transformer, homography estimation, image alignment, geometry-enhanced.**

I. Introduction

HOMOGRAPHY is a 3×3 matrix that contains 8 degree-of-freedoms (DOFs). It models the geometric transformation between two images of the planar surface captured from different perspectives. In non-coplanar scenarios, homography is often employed as an initial alignment model to establish a rough geometric transformation before applying more sophisticated techniques, such as mesh flow [1], [2] and optical flow [3], [4]. Homography estimation is crucial for various applications, including image alignment [5], [6], [7], video stabilization [8], [9], panoramic photography [10], [11], and camera calibration [12]. However, it is a challenge to estimate the homography of image pairs accurately, especially in cross-modal and cross-resolution scenarios.

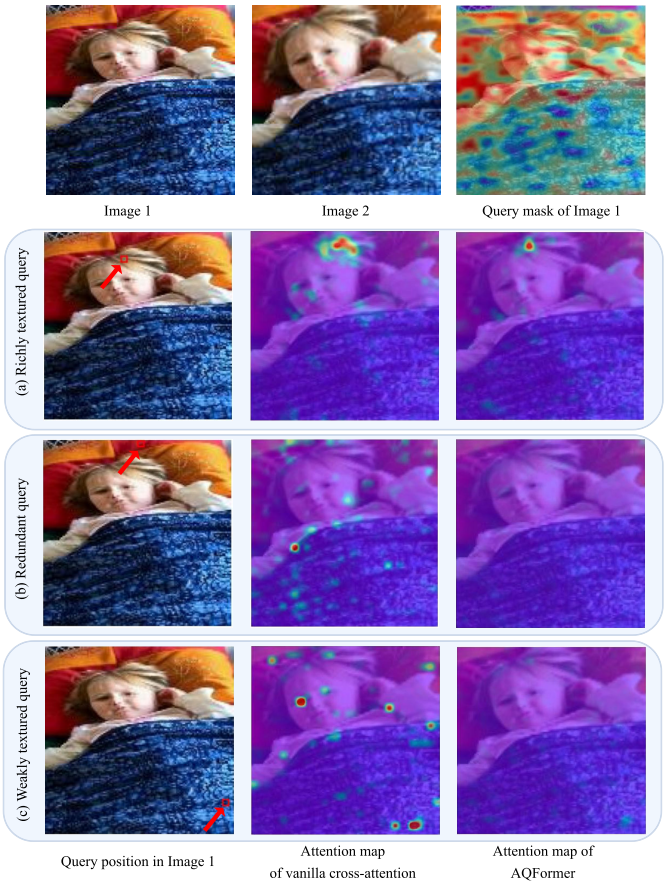


Fig. 1. Illustration of our main idea. We predict the query masks based on input images. (a), (b), and (c) visualize queries located in richly textured, redundant, and weakly textured areas, respectively, and the corresponding attention maps for these queries.

Since homography describes the overall geometric transformation relationship between image pairs, long-range dependency is pivotal for homography estimation. Recent methods employ vanilla cross-attention to establish correspondences across image pairs [13], [14]. Specifically, the model receives two sets of features: queries and key-value pairs. The purpose is to use one set of features (typically the query from Image 1) to selectively focus on relevant information in the other set (key-value pairs from Image 2), enabling the model to associate specific features across inputs. As shown in Fig. 1 (a), for the query of image 1, vanilla cross-attention efficiently focuses on the corresponding area of image 2.

Received 19 July 2024; revised 4 November 2024; accepted 12 November 2024. Date of publication 19 November 2024; date of current version 7 April 2025. This work was supported in part by the National Key Research and Development Program of China under Grant 2022ZD0161800 and in part by the National Natural Science Foundation of China under Grant 62271203. This article was recommended by Associate Editor B. Wen. *(Corresponding author: Faming Fang.)*

The authors are with the Department of Computer Science and Technology, East China Normal University, Shanghai 200062, China (e-mail: 52265901036@stu.ecnu.edu.cn; fmfang@cs.ecnu.edu.cn; tingtingwang@cs.ecnu.edu.cn; gxzhang@cs.ecnu.edu.cn).

Digital Object Identifier 10.1109/TCSVT.2024.3502170

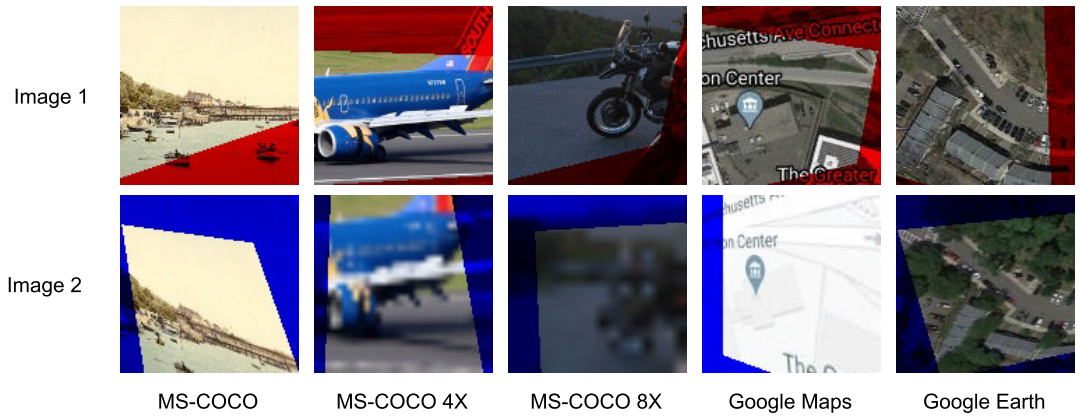


Fig. 2. Visualization of Input Image Pairs from Five Datasets. The unique content of Image 1 and Image 2 is shown in the R and B channels, respectively, while the shared content of the input image pairs is shown in RGB.

However, Vanilla cross-attention treats all queries uniformly which is suboptimal for homography estimation. Fig. 2 shows that estimating geometric transformations requires aligning shared regions between the images. Content unique to each image does not contribute to homography estimation which is termed "redundant areas". As shown in Fig. 1 (b), when unique areas appear in one image but not in the other, vanilla cross-attention focuses on incorrect regions. Moreover, as illustrated in Fig. 1 (c), it is challenging to establish accurate correspondences when the query is situated in weakly textured areas. Vanilla cross-attention treats queries in redundant and low-texture areas the same as those in richly textured areas, leading to the accumulation and propagation of erroneous information. We define this phenomenon, where the model excessively attends to queries in redundant and low-texture areas, as *query over-focusing*.

To address the issues, we introduce an **A**daptive **Q**uery trans**Former** termed AQFormer. Specifically, we design a **MA**sk **G**eneration module (MAG) that generates a mask based on the input image pairs. The mask corresponds to redundant and weakly textured areas with low weights, while it has high weights for richly textured areas. As shown in Fig. 1, the child and the pillow exhibit rich textures, playing a crucial role. The quilt with repetitive textures introduces interference in homography estimation. MAG generates a mask with higher weights in the regions of the child and the pillow, and lower weights in the quilt area. Additionally, MAG assigns lower weights for the area which only present in image 1. By applying this mask to the queries before cross-attention, AQFormer adaptively focuses on the most information-dense regions, effectively suppressing less relevant areas. Compared with vanilla cross transformers, AQformer brings about a notable enhancement, with the added parameters and computational complexity almost negligible.

The global information captured by AQFormer is fed into a homography aggregator to generate a coarse homography for the overall alignment. However, there are still some discrepancies in local details. We introduced the **G**ated **I**nteraction **M**odule (GIM) to refine the local details further. Specifically, GIM leverages the feature to generate gated signals which are then modulated into the convolutional kernel. Subsequently, the modulated kernel is applied to a conventional convolution operation to capture local information. Based on the gating mechanism, GIM focuses on relevant features and filters out unimportant information between feature pairs.

Integrating **A**QFormer and **G**IM into a multi-scale iterative framework, we propose a novel homography estimation network, termed **AGNet**. AGNet achieves state-of-the-art performance on five benchmark datasets, including challenging scenarios such as cross-resolution and cross-modal scenarios. The main contributions are summarized as follows:

- We discover the phenomenon of "query over-focusing", which interferes with the accuracy of cross-attention. The phenomenon is common in image pairs where the perspective changes.
- We propose an adaptive query transformer called AQFormer to mitigate query over-focusing. AQFormer suppresses the queries located in redundant and weakly textured areas and promotes the queries located in richly textured areas by an adaptive mask.
- We propose a gated interaction module called GIM. The GIM selectively captures local information by adjusting convolutional kernels based on input, enhancing the extraction of shared features between image pairs.

II. Related Work

A. Feature-Based Homography Estimation

The feature-based homography estimation methods typically work in four steps: key point detection, local feature extraction for each key point, feature-based key point matching, and matching-based homography fitting [15]. It can be seen that accurate matching is the key to traditional homography estimation.

Numerous studies have explored traditional key point matching methods [16], [17]. These methods initially detect key points based on the image information such as gradient. Subsequently, they calculate descriptors using the key points and the surrounding information. Thereafter, the descriptors are compared using metrics such as Euclidean distance to obtain the matched pairs. Finally, RANSAC [18] and MAGSAC [19] are used to remove outlier matches.

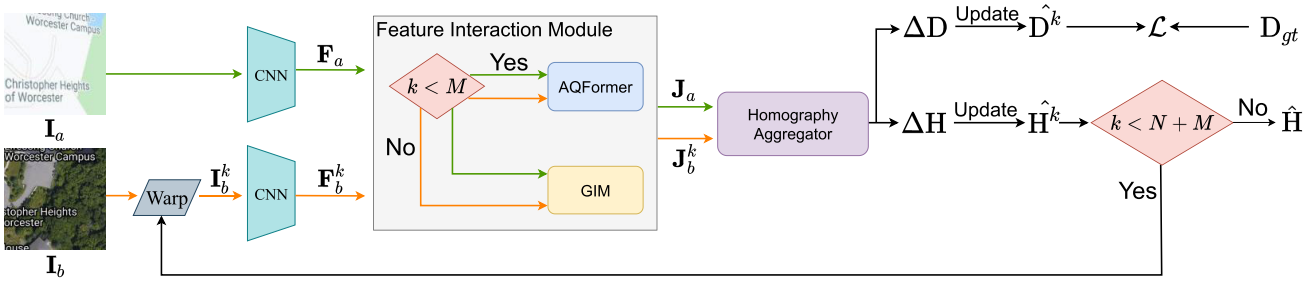


Fig. 3. The architecture of the proposed AGNet. $\mathbf{I}_a$ and $\mathbf{I}_b$ denote the input image pair. In the k-th iteration, $\mathbf{I}_b$ is warped into $\mathbf{I}_b^k$ guided by $\hat{\mathrm{H}}^{k-1}$. $\mathbf{F}_a$ and $\mathbf{F}_b^k$ are shallow features extracted by a Siamese network. $\mathbf{J}_a$ and $\mathbf{J}_b^k$ are fused features which are then sent to an aggregator to estimate ΔH and ΔD. ΔH and ΔD denote the homography and displacement vectors of the 4 corner points for aligning $\mathbf{I}_b^k$ to $\mathbf{I}_a$, respectively. $\hat{\mathrm{D}}^{k-1}$ and $\hat{\mathrm{H}}^{k-1}$ is updated by ΔD and ΔH to generate $\hat{\mathrm{D}}^k$ and $\hat{\mathrm{H}}^k$, which are then used to supervise the network learning and warp $\mathbf{I}_b$, respectively. After $M+N$ iterations, AGNet outputs the final homography $\hat{\mathrm{H}}$.

Some methods introduce richer structural information, such as line segments, to extract more features [20], [21]. However, insufficient feature extraction remains a pain point for traditional methods. Traditional methods rely on manually crafted descriptors, which may struggle to capture complex representations in diverse visual data, such as cross-modal and cross-resolution scenarios [22], [23], [24].

Deep learning methods [25], [26], [27] gain prominence in key point matching. SuperPoint [25] leverages a fully convolutional architecture to generate dense feature maps. Additionally, SuperGlue [26] performs feature matching and geometric verification, improving the accuracy and reliability of correspondences. LoFTR [27] employs a transformer architecture, showcasing superior performance in challenging scenarios. Compared to traditional matching methods, deep matching methods are more accurate. However, these feature-based homography estimation methods involve multiple stages. The complex pipeline increases computational costs and processing time.

B. CNN-Based Homography Estimation

In recent years, CNN-based homography estimation has undergone significant advancements. As a pioneering work, DHN [28] introduces a CNN architecture that directly outputs homography matrices based on input image pairs. UDHN [29] enhances this by proposing a pixel-wise photometric loss for unsupervised training. Additionally, several works [1], [30], [31] have demonstrated strong performance in real-world scenarios using pixel-wise photometric loss. ECLUH [32] incorporates intuitive structural information as an additional cue, making it more sensitive to human vision and effective in low-texture situations. Nie et al. [33] design a contextual correlation layer to extract the matching relationships from global to local. However, these unsupervised methods exhibit instability and convergence challenges during training [34]. Their applicability is constrained to image pairs with small baselines, such as continuous video frames or photos captured by dual-camera smartphones.

To estimate homography with large baselines, CLKN [35] learns deep features through a recurrent framework and employs an inverse compositional algorithm based on iterative closest Lucas Kanade (IC-LK). Furthermore, DLKFM [36] extends the applicability of IC-LK to estimate homography for cross-modal image pairs by introducing a novel loss function. However, IC-LK is an untrainable layer of the deep network. IHN [37] abandon untrainable IC-LK and design an iterative architecture that can be trained end-to-end, predicting the homography from coarse to fine, significantly improving prediction accuracy.

C. Transformer-Based Homography Estimation

Transformers and their variants have been applied to homography estimation in recent years. HomoGAN [38] proposes an unsupervised GAN using the transformer architecture as the backbone to impose coplanarity constraints on the predicted homography. LocalTrans [14] designs a cross-transformer module to capture the long-short range dependencies between image pairs. Following this, RHWF [13] indicates that standard convolutions fail to uphold equivariance beyond translation. Consequently, they utilize homography to warp images instead of feature maps. However, the above methods overlook query over-focusing in homography estimation. Our approach adaptively adjusts the weight of each query, focusing on prominent features in the shared regions of image pairs.

III. Methodology

The architecture of our proposed AGNet for homography estimation is shown in Fig. 3, which can be divided into three components: feature extraction, feature interaction by adaptive query transformer (AQFormer) or gated interaction module (GIM), and homography aggregator.

A. Overview

The AGNet takes two images $\mathbf{I}_a \in \mathbb{R}^{H\times W\times 3}$ and $\mathbf{I}_b \in \mathbb{R}^{H\times W\times 3}$ as input and outputs a homography $\hat{\mathrm{H}}$ that aligns $\mathbf{I}_b$ to $\mathbf{I}_a$, where H and W denote the height and width of the image. To enhance the accuracy of homography estimation, we refine the homography iteratively. In the k-th iteration, $\mathbf{I}_b$ is warped to $\mathbf{I}_b^k$ guided by $\hat{\mathrm{H}}^{k-1}$, achieving a coarse alignment with $\mathbf{I}_a$. Note that $\hat{\mathrm{H}}^0$ is initialized as the identity matrix. Then, $\mathbf{I}_a$ and $\mathbf{I}_b^k$ are fed into a Siamese network to obtain the features

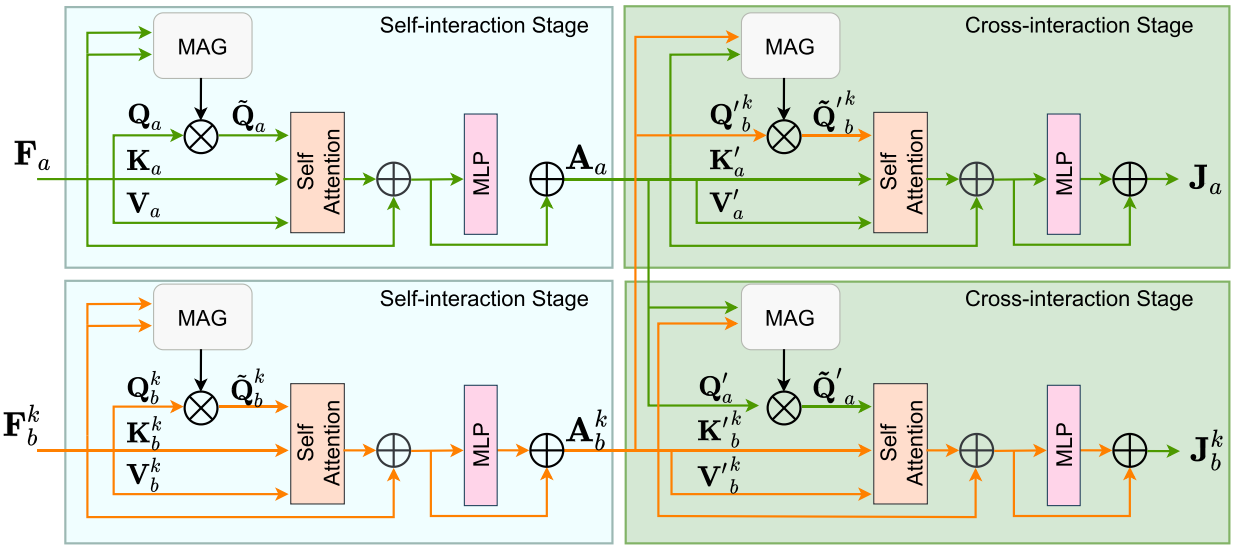


Fig. 4. The architecture of the adaptive query transformer (AQFormer) that captures global dependencies.

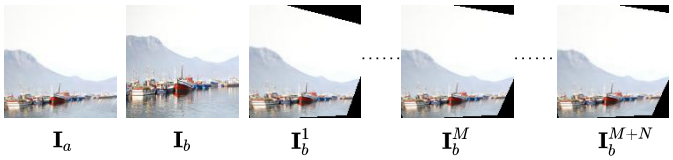


Fig. 5. Illustration of warping $\mathbf{I}_b$ under the guidance of $\hat{\mathrm{H}}^k$.

$\mathbf{F}_a \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C}$ and $\mathbf{F}_b^k \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C}$, where C denotes the channel dimension. To establish correspondences, we take $\mathbf{F}_a$ and $\mathbf{F}_b^k$ as input into the feature interaction module (FIM). Based on the output of the FIM, a homography aggregator estimates ΔH and ΔD which denote the homography and displacement vectors of the 4 corner points for aligning $\mathbf{I}_b^k$ to $\mathbf{I}_a$, respectively. Finally, $\hat{\mathrm{D}}^{k-1}$ is updated by ΔD to generate $\hat{\mathrm{D}}^k$, which is used to supervise the network learning. $\hat{\mathrm{H}}^{k-1}$ is updated by ΔH to generate $\hat{\mathrm{H}}^k$, which is used to warp $\mathbf{I}_b$ in the $\{k+1\}$-th iteration. After $M+N$ iterations, AGNet outputs the final homography $\hat{\mathrm{H}}$, where M and N are hyperparameters controlling the number of iterations.

B. Feature Interaction Module

Deep feature interaction between two images plays a crucial role in homography estimation. It matches corresponding points or regions across the images, which is essential for accurately estimating the transformation. We propose a feature interaction module (FIM) to facilitate feature interaction from global to local levels, refining the homography estimation progressively. The FIM sends $\mathbf{F}_a$ and $\mathbf{F}_b^k$ to AQFormer or GIM based on the number of iterations and outputs the fused features $\mathbf{J}_a$ and $\mathbf{J}_b^k$. As shown in Fig. 5, significant geometric disparities exist between the image pairs in the early iterations. In the later iterations, $\mathbf{I}_b$ is approximately aligned with $\mathbf{I}_a$ under the guidance of $\hat{\mathrm{H}}^{k-1}$. Therefore, $\mathbf{F}_a$ and $\mathbf{F}_b^k$ are fed to AQFormer during the first M iterations. In the subsequent N iterations, $\mathbf{F}_a$ and $\mathbf{F}_b^k$ are sent to the GIM, which specializes in capturing local information.

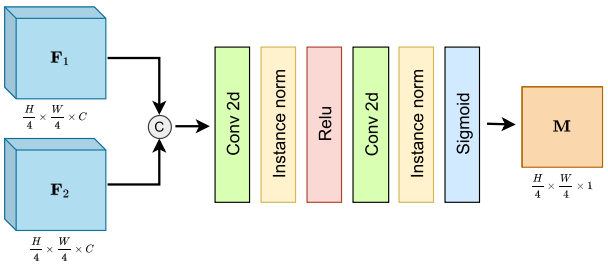


Fig. 6. The detailed architecture of the mask generation module (MAG). $\mathbf{F}_1$ denotes the feature that is subsequently mapped as the query, $\mathbf{F}_2$ denotes the feature that is subsequently mapped as the key and value, and $\mathbf{M}$ denotes the mask applied to the queries.

C. Adaptive Query Transformer

Recent transformer-based methods [13], [14] demonstrate that global information is profitable for capturing the geometric transformation between image pairs. However, the vanilla attention map may be misdirected when the keys respond strongly to queries in redundant or weakly textured areas. This inaccurate attention map then influences the values, leading to incorrect global correspondences. Hence, we design an adaptive query transformer that adaptively adjusts the weights of queries in different regions based on input features.

To achieve the above objectives, we develop a CNN-based mask generation module termed MAG. Formally, the feature subsequently mapped as the query is denoted as $\mathbf{F}_1$, and the feature subsequently mapped as the key and value is denoted as $\mathbf{F}_2$. As shown in Figure 5, we concatenate $\mathbf{F}_1$ and $\mathbf{F}_2$ and feed them into a two-layer convolutional network to generate a mask $\mathbf{M}$ with a channel dimension of 1, which is used to adjust the weights of the queries. Specifically, queries located in rich texture areas are assigned higher weights, while those located in redundant or weakly textured areas are assigned weights close to 0. The process can be expressed as:

$$\mathbf{M} = \mathcal{M}(CAT(\mathbf{F}_1, \mathbf{F}_2)), \tag{1}$$

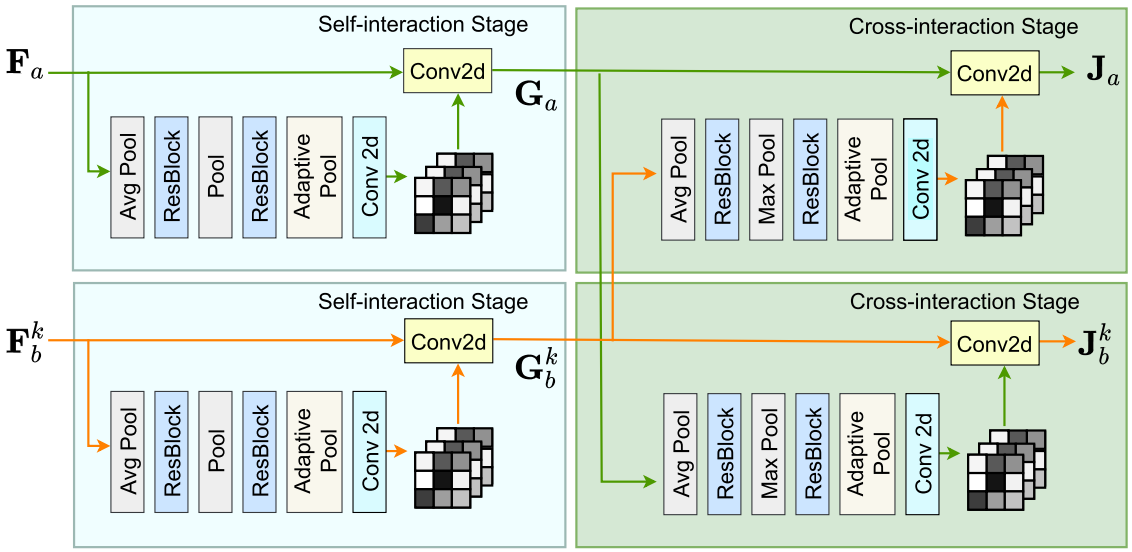


Fig. 7. The architecture of the gated interaction module (GIM). **F** denotes the input pair, **G** denotes the middle feature, and **J** denotes the output of GIM.

where $CAT(\cdot)$ denotes concatenation operation along the channel dimension, and $\mathcal{M}$ denotes the mask generation network.

Combining with MAG, AQFormer consists of two stages: self-interaction and cross-interaction, which is illustrated in Fig. 4. The shallow features $\mathbf{F}_a$ and $\mathbf{F}_b^k$ are first fed into a self-interaction encoder. In this stage, each feature map undergoes self-interaction to generate deep features, denoted as $\mathbf{A}_a$ and $\mathbf{A}_b^k$, respectively. Subsequently, $\mathbf{A}_a$ and $\mathbf{A}_b^k$ are then fed into a cross-interaction encoder. In this stage, $\mathbf{A}_a$ and $\mathbf{A}_b^k$ interact with one another to merge their information and generate the final feature $\mathbf{J}_a$ and $\mathbf{J}_b^k$.

1) Self-Interaction of AQFormer: We first adopt three 1×1 convolution layers $f_Q(\cdot)$, $f_K(\cdot)$, and $f_V(\cdot)$ to encode the input image feature $\mathbf{F}$ ($\mathbf{F}_a$ or $\mathbf{F}_b^k$) to the features $\mathbf{Q}, \mathbf{K}$ and $\mathbf{V}$. The process is expressed as follows:

$$\mathbf{Q}, \mathbf{K}, \mathbf{V} = f_Q(\mathbf{F}), f_K(\mathbf{F}), f_V(\mathbf{F}). \tag{2}$$

In self-attention, even though the features for generating $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ are identical, we leverage MAG to decrease the weight of queries in areas of weak texture. The process is expressed as follows:

$$\tilde{\mathbf{Q}} = \mathcal{M}(CAT(\mathbf{F}, \mathbf{F})) \odot \mathbf{Q}, \tag{3}$$

where $\tilde{\mathbf{Q}}$ denotes the adaptive query of $\mathbf{Q}$ and $\odot$ denotes the element-wise product operation. Then the self-attention can be formulated as:

$$\mathbf{A} = \mathcal{S}(\frac{\tilde{\mathbf{Q}}\mathbf{K}}{\sqrt{C}})\mathbf{V}, \tag{4}$$

where $\mathcal{S}$ denotes the SoftMax function.

2) Cross-Interaction of AQFormer: Since $\Theta(\mathbf{A}_a, \mathbf{A}_b^k)$ and $\Theta(\mathbf{A}_b^k, \mathbf{A}_a)$ are similar, where Θ represents cross-interaction. For simplicity, we only elaborate on $\Theta(\mathbf{A}_a, \mathbf{A}_b^k)$ in the following text. We first adopt two 1×1 convolution layers $f'_Q(\cdot)$ to encode the deep feature $\mathbf{A}_b^k$ to the features $\mathbf{Q}'^k_b$. Then, the convolutional layer $f'_K(\cdot)$ and $f'_V(\cdot)$, which shares the same architecture as $f'_Q(\cdot)$ but does not share weights, is utilized to encode the deep feature $\mathbf{A}_a$ to generate the features $\mathbf{K}'_a$ and $\mathbf{V}'_a$. The process is expressed as follows:

$$\begin{aligned} \mathbf{Q}'^k_b &= f'_Q(\mathbf{A}_b^k), \\ \mathbf{K}'_a, \mathbf{V}'_a &= f'_K(\mathbf{A}_a), f'_V(\mathbf{A}_a). \end{aligned} \tag{5}$$

Due to significant geometric transformations in $\mathbf{A}_a$ and $\mathbf{A}_b^k$, some queries in $\mathbf{Q}'^k_b$ cannot match correct correspondences in $\mathbf{K}'_a$. Directly computing the attention matrix using $\mathbf{Q}'^k_b$ and $\mathbf{K}'_a$ is not suitable. Therefore, we utilize MAG to decrease the weights of queries in redundant and weakly textured areas as follows:

$$\tilde{\mathbf{Q}}'^k_b = \mathcal{M}(CAT(\mathbf{A}_b^k, \mathbf{A}_a)) \odot \mathbf{Q}'^k_b, \tag{6}$$

where $\tilde{\mathbf{Q}}'^k_b$ denotes the adaptive query of $\mathbf{Q}'^k_b$. The cross-attention can be formulated as:

$$\mathbf{J}_a = \mathcal{S}(\frac{\tilde{\mathbf{Q}}'^k_b \mathbf{K}'_a}{\sqrt{C}})\mathbf{V}'_a. \tag{7}$$

Similar to the vanilla attention architecture, The residual connecting, layer normalization, and Multilayer Perceptron (MLP) are applied at both self-interaction and cross-interaction stages to obtain the final results.

D. Gated Interaction Module

AGNet outputs $\hat{\mathrm{H}}^M$ in M iterations. Guided by $\hat{\mathrm{H}}^M$, $\mathbf{I}_b$ is warped to $\mathbf{I}_b^M$, which is roughly aligned with $\mathbf{I}_a$. As shown in Fig. 5, there are only minor local misalignments between $\mathbf{I}_a$ and $\mathbf{I}_b^M$. It is unnecessary to establish global dependencies. Consequently, we abandon the transformer and utilize the CNN as the backbone for subsequent iterations. Inspired by gMLP [39], we introduce a Gated Interaction Module (GIM) as shown in Fig. 7. The GIM adjusts the convolutional kernels flexibly based on the input information, allowing it to extract common information from the input pair.

Like AQFormer, GIM is divided into two stages: feature self-interaction and feature cross-interaction. The shallow features $\mathbf{F}_a$ and $\mathbf{F}_b^k$ are initially fed into the feature self-interaction

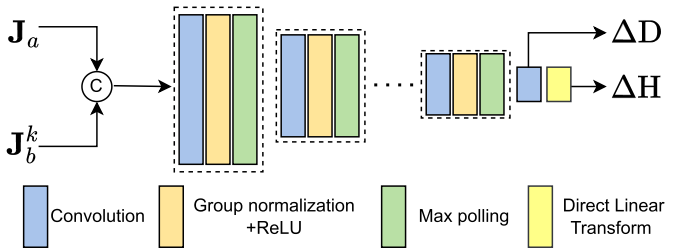


Fig. 8. The detailed architecture of the homography aggregator.

encoder to produce deep features $\mathbf{G}_a$ and $\mathbf{G}_b^k$. Then, $\mathbf{G}_a$ and $\mathbf{G}_b^k$ are inputted into the cross-interaction encoder to generate the final feature $\mathbf{J}_a$ and $\mathbf{J}_b^k$.

1) Self-Interaction of GIM: We design a gating network that generates C convolutional kernels with the spatial size of 3×3 pixels based on the input feature $\mathbf{F}$ ($\mathbf{F}_a$ or $\mathbf{F}_b^k$) $\in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C}$. Specifically, we first apply pooling followed by convolution layer twice, to generate features with spatial dimensions of $\frac{H}{16} \times \frac{W}{16}$ and channel dimension of C. Next, we use adaptive pooling to compress the spatial dimensions of the feature to 1×1, and then apply a convolution layer to generate a feature of size $1 \times 1 \times (C \times 3 \times 3)$. Finally, we reshape the feature into C convolutional kernels with the spatial size of 3×3 pixels. These kernels are then applied to $\mathbf{F}$ to extract features. The process can be expressed as:

$$\mathbf{G} = \mathcal{N}_{self}(\mathbf{F}) \otimes \mathbf{F}, \tag{8}$$

where $\mathcal{N}_{self}$ denotes the gating network, $\mathbf{G}$ ($\mathbf{G}_a$ or $\mathbf{G}_b^k$) denotes the deep feature produced by self-interaction of GIM and $\otimes$ denotes a convolution operation.

2) Cross-Interaction of GIM: $\mathbf{G}_a$ and $\mathbf{G}_b^k$ originate from two different images, which may even have different resolutions and modalities. Concatenating $\mathbf{G}_a$ and $\mathbf{G}_b^k$ directly to estimate the homography is not elegant. Although $\mathbf{G}_a$ and $\mathbf{G}_b^k$ have distributional differences, they have shared features. We discard the unique features of $\mathbf{G}_a$ and $\mathbf{G}_b^k$, only utilizing their shared features.

Since $\Gamma(\mathbf{G}_a, \mathbf{G}_b^k)$ and $\Gamma(\mathbf{G}_b^k, \mathbf{G}_a)$ are similar, where Γ represents cross-interaction of GIM. For simplicity, we only elaborate on $\Gamma(\mathbf{G}_a, \mathbf{G}_b^k)$ in the following text. We first feed $\mathbf{G}_b^k$ into a gated network $\mathcal{N}_{cross}$ to generate C convolutional kernels with the spatial size of 3×3 pixels. Next, these convolutional kernels are applied to $\mathbf{G}_a$ to extract the shared features with $\mathbf{G}_b^k$, while discarding the unique features of $\mathbf{G}_a$. The process can be expressed as:

$$\mathbf{J}_a = \mathcal{N}_{cross}(\mathbf{G}_b^k) \otimes \mathbf{G}_a, \tag{9}$$

where $\mathcal{N}_{cross}$ denotes the gating network with the same structure as $\mathcal{N}_{self}$, but with different weights.

E. Homography Aggregator

The homography aggregator estimates the homography that aligns $\mathbf{I}_b^k$ to $\mathbf{I}_a$ based on the output of the FIM. Traditional computer graphics typically treat homography as a 3×3 matrix with 8 degrees of freedom. However, supervising homography presents challenges during training because the homography matrix combines rotation and translation terms, which are difficult to balance. To address this, we estimate the displacement of four corner points instead of directly supervising the homography matrix. Once the displacement of the four corners is known, we use the normalized Direct Linear Transform (DLT) algorithm to calculate the homography matrix [28].

The detailed architecture of the homography aggregator is shown in Fig. 8, $\mathbf{J}_a$ and $\mathbf{J}_b^k$ are concatenated and then sent to the aggregator which consists of multiple basic units. Each unit includes a 3×3 convolution, followed by group normalization and ReLU activation, and a max-pooling layer with a stride of 2. By stacking basic units, a feature map with a spatial resolution of 2×2 is obtained. Subsequently, a convolutional layer projects the feature map into a $2 \times 2 \times 2$ cube ΔD. ΔD denotes the displacement vectors of the 4 corner points for aligning $\mathbf{I}_b^k$ to $\mathbf{I}_a$. ΔD is transformed into an equivalent ΔH by DLT algorithm. Clearly, $\hat{\mathrm{D}}^k = \hat{\mathrm{D}}^{k-1} + \Delta\mathrm{D}$ and $\hat{\mathrm{H}}^k = \hat{\mathrm{H}}^{k-1}\Delta\mathrm{H}$.

F. Multiscale Refinement

Previous studies have indicated that multi-scale refinement further enhances the performance of networks. Taking inspiration from these findings [13], [14], [37], we design the 2-scale AGNet. Specifically, after $M + N$ iterations at low resolution, $\mathbf{I}_a$ and $\mathbf{I}_b^{M+N}$ are input into a new Siamese network to obtain the feature pairs with a resolution of $\frac{H}{2} \times \frac{W}{2}$. The feature pairs are then sent to the FIM for further feature processing. We iterate an additional K times at high resolution to achieve a more accurate refinement. The experiments demonstrate that the 1-scale AGNet outperforms the majority of prior works. The two-scale AGNet exhibits a significant performance improvement compared to the 1-scale AGNet.

IV. Experiment

In this section, we first describe the utilized datasets and the specific training configurations. Then, we present the implementation details. Finally, we compare our results with both feature-based and deep homography methods on five benchmark datasets for common scenarios, cross-resolution scenarios, and cross-modal scenarios.

A. Datasets

1) MS-COCO: The MS-COCO dataset is a large-scale real-world RGB dataset that has been widely used in recent deep homography estimation approaches. To generate the training pairs, a square patch, denoted as $\mathbf{I}_p$, is first cropped from a larger image $\mathbf{I}$ at a random position p. The four vertices of this patch are then randomly displaced within a range of [-ρ, ρ] pixels, where ρ represents the maximum allowable displacement, to form a new patch. Given that the degrees of freedom for a homography is 8, a homography matrix $\mathbf{H}_{AB}$ is computed based on the four corresponding points. The inverse homography $\mathbf{H}_{BA} = (\mathbf{H}_{AB})^{-1}$ is applied to the original image $\mathbf{I}$, producing a transformed image $\mathbf{I}'$. A second patch $\mathbf{I}'_p$ is cropped from $\mathbf{I}'$ at the same position p. As a result, $\mathbf{I}_p$ and $\mathbf{I}'_p$ form the input image pair, with $\mathbf{H}_{AB}$ serving as the ground truth. To evaluate the accuracy, we use the Average Corner Error (ACE) as the metric.

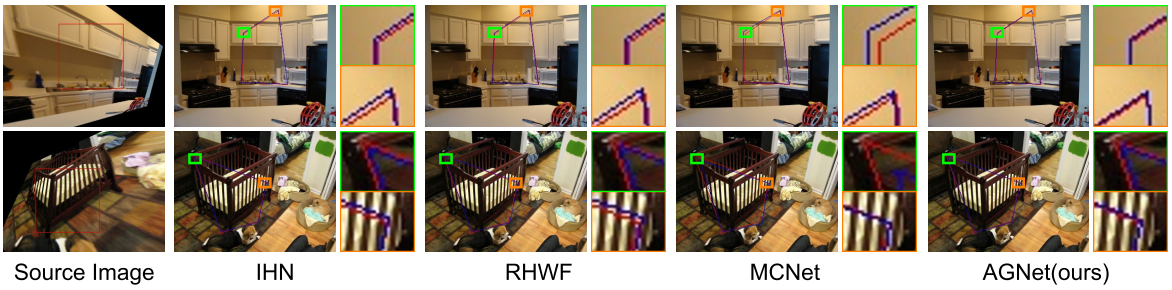


Fig. 9. Visualization of results on the MS-COCO dataset. The red polygon represents the ground position of the source image on the target image. The blue polygon represents the estimated position using different algorithms on the target image. The closer the two colors of polygons are, the better the estimation accuracy.

To demonstrate the capability of AGNet in cross-resolution scenarios, we perform downsampling on the target images using bicubic interpolation with scaling factors of 4× and 8×. For a more comprehensive comparison, we categorized the results into three levels based on ACE, defining the top 30% as Easy, the 30% − 70% range as Medium, and the bottom 30% as Hard.

2) Google Earth: Google Earth provides satellite images on different dates. Following the setup of DLKFM [36], images from April 2018 and June 2019 in the Greater Boston area are selected as cross-modal input data. Due to different shooting times, the image pairs exhibit variations in both structure and color. We employ a similar method to MS-COCO to introduce displacements. There are a total of 8,750 image pairs for training and 850 image pairs for testing.

3) Google Maps: Static Google maps and satellite maps can be obtained by the Google Maps API. They represent the same area with distinct color patterns. Following the DLKFM configuration [36], we have 8,822 cross-modal image pairs for training and 888 cross-modal image pairs for testing.

B. Implementation Details

We implement our method based on PyTorch 1.7.0 and train it from scratch using a machine with one NVIDIA GeForce RTX 3090 GPU. The batch size is set to 16. We use the AdamW optimizer [40] with default parameter settings as the optimizer. The learning rate is initialized to be 0.0004 and is updated by the OneCycleLR scheme. The iteration times M, N, and K are set to 3, 3, and 6. The hyperparameters γ and ρ are set to 0.85 and 32.

C. Evaluation on MS-COCO

We compared AGNet on the MS-COCO dataset with other homography estimation methods, including deep homography methods such as DHN [28], UDHN [29], LocalTrans [14], IHN [37], RHWF [13], ECLUH [32], MCNet [41] and feature-based homography methods such as SIFT+RANSAC [16], LoFTR [27], and GeoFormer [34].

Table I presents the quantitative evaluation results. In easy scenarios, feature-based methods capture the relationship between corresponding feature points to solve homography. However, in hard scenarios, these methods may match incorrect feature points. Consequently, the subsequent homography calculation exacerbate the error, leading to failure. In contrast, deep methods circumvent matching feature points and estimate homography end-to-end, which leads to more robust performance in complex scenarios. Compared to the previous state-of-the-art (SOTA) method RHWF [13], our method demonstrates a significant improvement in accuracy, achieving a 30% enhancement.

TABLE I

THE MACE COMPARISON ON THE MS-COCO DATASET. ROWS 1-6 REPRESENT DEEP HOMOGRAPHY METHODS, WHILE ROWS 7, 8, AND 9 CORRESPOND TO FEATURE-BASED HOMOGRAPHY METHODS

Methods	Easy	Medium	Hard	Average	Param(M)
DHN	3.5187	5.9704	11.197	6.8028	34.19
UDHN	0.5408	2.4335	11.952	4.7212	21.29
LocalTrans	0.0968	0.1500	0.5769	0.2621	9.57
IHN	0.0255	0.0452	0.1174	0.0610	1.71
1-scale RHWF	0.0254	0.0513	0.1445	0.0715	0.93
2-scale RHWF	0.0061	0.0120	0.0575	0.0239	1.29
ECLUH	0.5200	0.7200	2.3100	1.2300	-
MCNet	0.0968	0.1500	0.5769	0.2621	0.85
SIFT+RANSAC	0.2836	0.6986	74.467	22.692	N/A
LoFTR	0.4715	1.0174	7.9051	2.9201	11.56
GeoFormer	1.2471	2.6971	7.0942	3.5812	14.19
1-scale AGNet	0.0230	0.0463	0.1215	0.0619	1.56
2-scale AGNet	**0.0049**	**0.0101**	**0.0373**	**0.0167**	2.34

Fig. 9 illustrates the visualization results of aligning Image 2 to Image 1 using the estimated homography. Despite the sky occupying the majority of the image pairs, our method effectively eliminates the interference from sky-dominated textureless regions and aligns the kite, which contains rich visual information.

D. Evaluation on Cross-Resolution MSCOCO

For multi-scale gigabit pixel photography, cross-resolution homography evaluation is a crucial aspect [11]. Images captured by different sensors or at different scales often need to be processed together. Compared to common scenarios, the decrease in resolution greatly increases the difficulty of homography estimation. Following LocalTrans [14], we conduct the evaluation on 4× and 8× cross-resolution MSCOCO.

The visualization results are shown in Fig. 11. In the first case, both IHN and RHWF estimate inaccurate homography due to the low resolution. Only our method preserves the tail of the animal after alignment. In the second case, due to the low resolution, the two lines in Image 1 degenerate into a single black line in Image 2. Both IHN and RHWF incorrectly align the black line in Image 2 to the bottom line in Image 1.

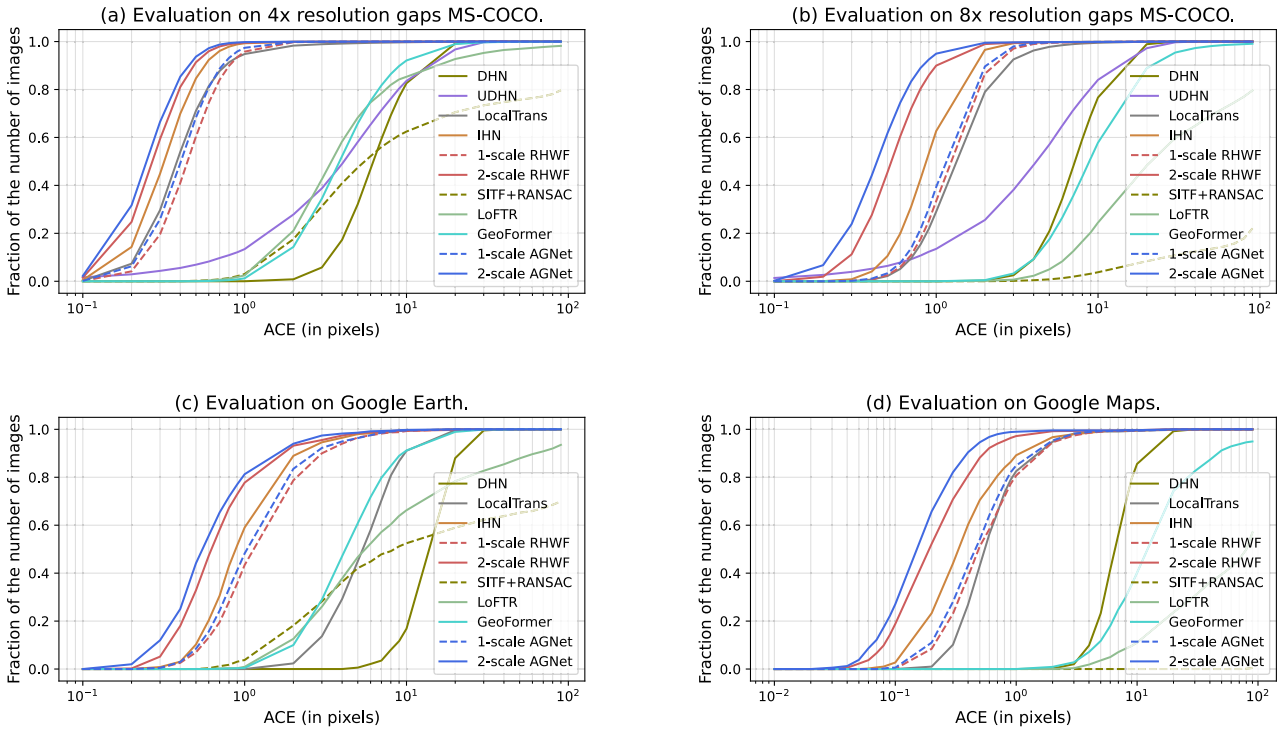


Fig. 10. Homography estimation evaluation. The x-axis is the mean average pixel error, the y-axis is the cumulative percentage of test images that have a lower average pixel error than x.

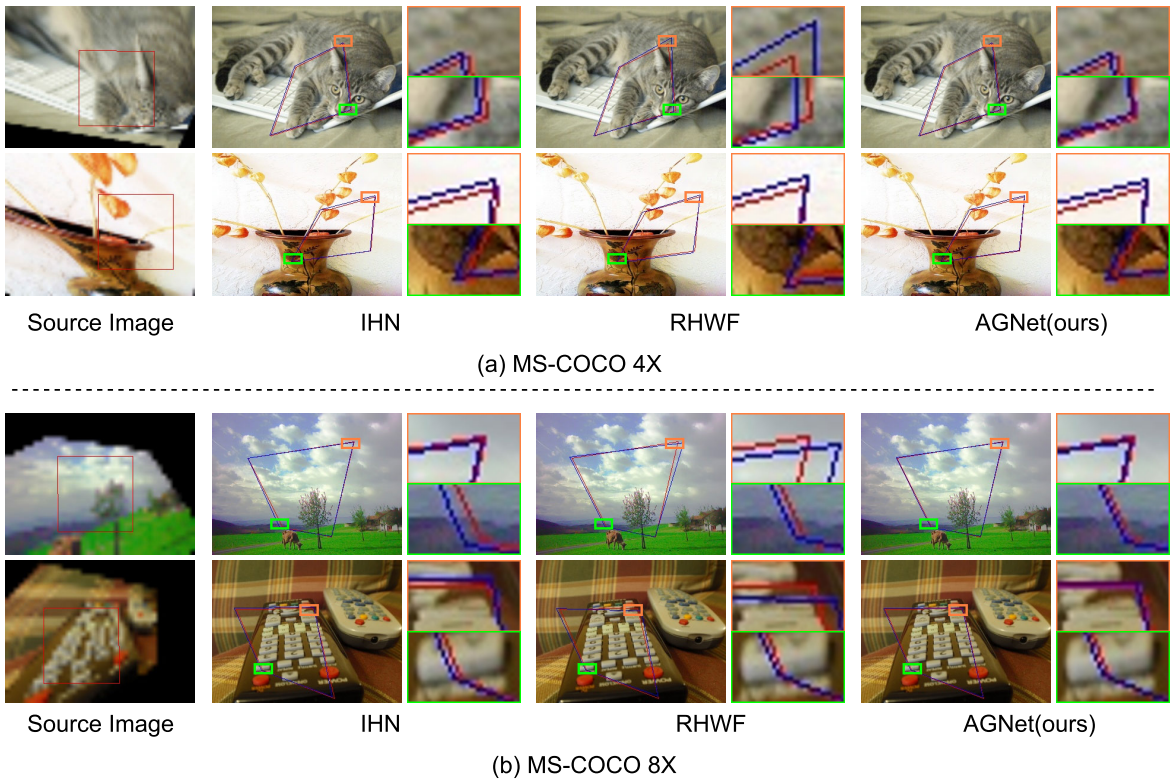


Fig. 11. Visualization of results on the MS-COCO 4× and 8× dataset.

However, our method warp the black line to a more accurate position.

Quantitative comparison is illustrated in Fig. 10(a) and Fig. 10(b). we plot the fraction of the number of images with respect to the corresponding ACEs of the dataset. Feature-based methods struggle in cross-resolution scenes due to the necessity of extracting features from two images and subsequently matching them. By obtaining a coarse estimation of the homography, our method establishes a preliminary alignment between the images, regardless of their resolution

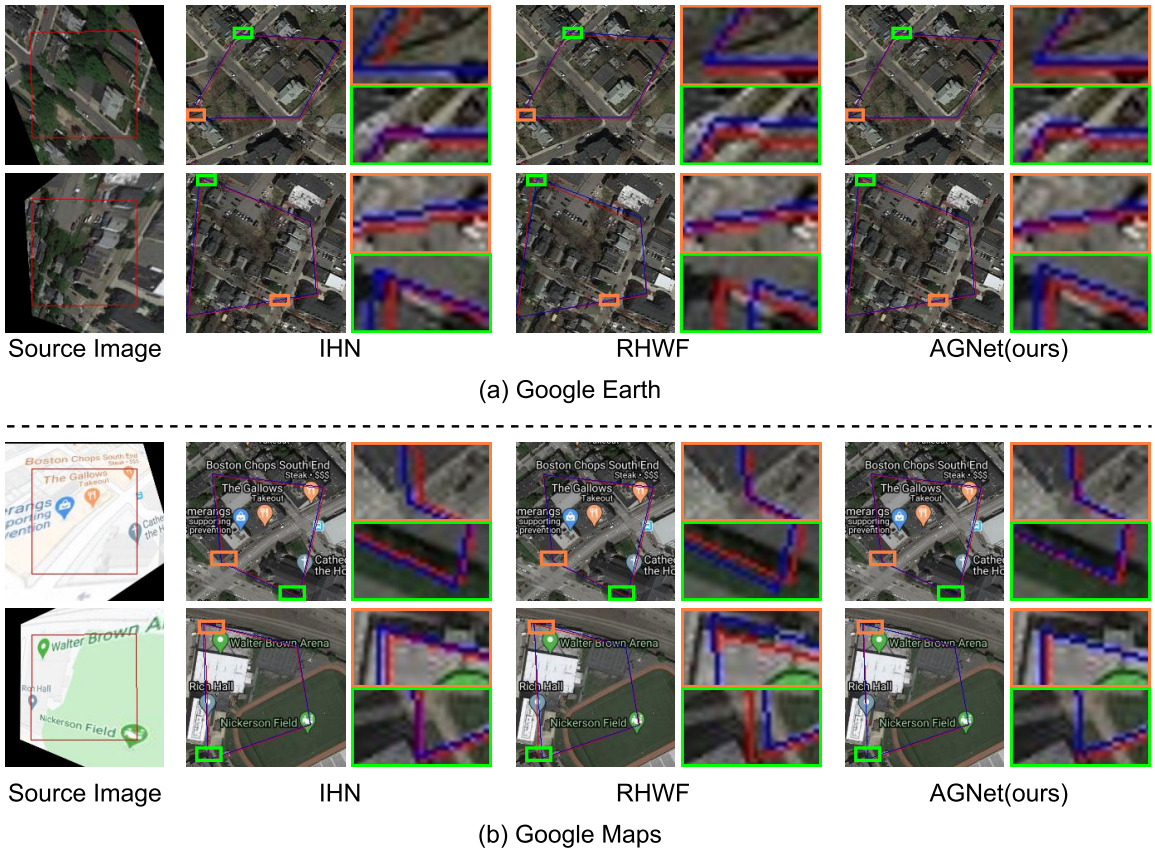


Fig. 12. Visualization of results on the Google Earth and Google Maps datasets.

disparities. This coarse estimation serves as a robust starting point, enabling subsequent fine-tuning steps to iteratively improve the accuracy of the homography. As a result, our method achieves superior performance in both MS-COCO 4× and MS-COCO 8× datasets.

E. Evaluation on Cross-Modal Datasets

In some scenarios, data from a specific modality may be unavailable. Cross-modal models compensate for missing data by utilizing information from other accessible modalities. However, Different modalities of data are typically captured using different devices. Even slight changes in angle or position of the devices can result in significant variations in the content. Additionally, data is sometimes collected at different times, and even if the capturing devices remain unchanged, the content being captured may still undergo alterations. It is an essential step to align image pairs from disparate modalities before conducting downstream tasks.

Different modalities of data have different distributions and representations. Google Earth encompasses images captured at the same geographic location but during different seasons. Google Maps provides satellite images and their corresponding maps. Effectively estimating the homography of cross-modal data poses a challenge. We evaluate the proposed models against the state-of-the-art methods on Google Earth and Google Maps datasets. The results are illustrated in Fig. 10(c) and Fig. 10(d). One can see that AGNet has obvious advantages over both deep homography methods and feature-based homography methods.

Fig. 12 shows the visualization results. For Google Earth datasets, IHN and RHWF exhibit obvious gaps. Our results are closer to the ground truth. For Google Maps datasets, IHN and RHWF produce structure inconsistency in the letter 's'. In contrast, our method creates an artifacts-free result.

V. Analysis and Discussions

In this section, we conduct ablation experiments to evaluate each module of AGNet comprehensively. All experiments in this section are executed on 1-scale AGNet with iteration times M and N both being 3, unless explicitly stated. To ensure a fair comparison, all variants are trained and evaluated under identical settings, encompassing training strategy, training equipment, data preprocessing, hyperparameters, and evaluation metrics.

A. Effectiveness of MAG

MAG predicts a mask based on the input feature pairs. The mask dynamically adjusts the weights of the queries. We conduct ablation experiments to assess its impact on homography estimation by removing MAG while keeping other settings unchanged. It's worth noting that after removing MAG, AQFormer degenerates into a vanilla transformer. As shown in Table II, MAG introduces more accurate homography with lower MACE values at all three levels. To further

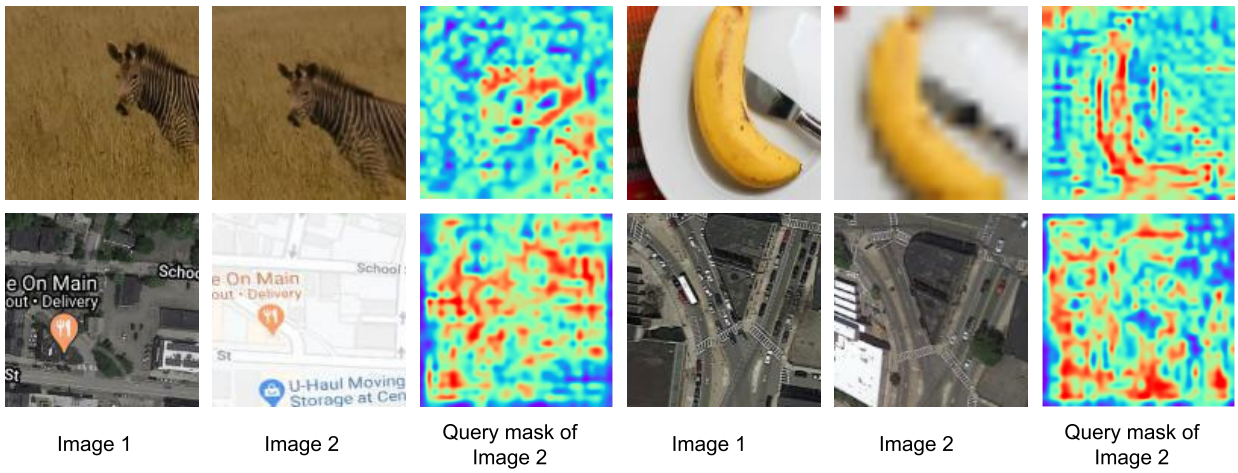


Fig. 13. Visualization results of masks generated by MAG.

TABLE II
THE MACE COMPARISON OF ABLATION STUDY ON THE MS-COCO DATASET

Methods	Easy	Medium	Hard	Average
w/o MAG	0.0244	0.0483	0.1302	0.0657
w/o AQFormer	**0.0230**	0.0488	0.1364	0.0673
w/o GIM	0.0265	0.0474	0.1266	0.0649
1-scale AGNet	**0.0230**	**0.0463**	**0.1215**	**0.0619**

TABLE III
THE MACE COMPARISON OF ABLATION STUDY ON THE LARGE BASELINE

Methods	Easy	Medium	Hard	Average
w/o AQFormer	0.0546	0.1548	10.920	3.3542
w/o GIM	0.0632	0.1622	8.2671	2.5640
1-scale AGNet	0.0514	0.1334	8.6946	2.6772
1-scale AGNet+	**0.0511**	**0.1291**	**7.1425**	**2.2097**

illustrate the importance of MAG, we present the visualization results of the masks generated by MAG. As shown in Fig. 13, whether in single-modal or cross-modal scenarios, the masks generated by MAG assign lower weights to redundant or weakly textured areas and high weights to richly textured areas.

B. Effectiveness of GIM

To demonstrate the effectiveness of GIM, we introduce a new variant (w/o GIM) by setting the iteration counts M and N to 6 and 0, respectively, The variant 'w/o GIM' ensures that AGNet only uses AQFormer for global information interaction without GIM for local information interaction. Table II shows that GIM improves the accuracy of homography estimation.

C. Effectiveness of AQFormer

We propose an AQFormer to better capture global information. To demonstrate the effectiveness of this design, we design a new variant (w/o AQFormer) that sets the iteration counts M and N to 0 and 6, respectively. The variant 'w/o AQFormer' ensures that AGNet only builds local dependencies using GIM without involving AQFormer for global dependencies. Table II presents the quantitative evaluation. We observe that discarding AQFormer results in an increase in the MACE by 0.0054. In hard scenarios, discarding AQFormer leads to even greater performance degradation. This indicates the importance of global information in homography estimation, particularly for large baselines. In the following subsection, we discuss global information's impact on large baselines.

D. Study of Large Baselines

To investigate the effectiveness of AQFormer in large baseline scenarios where image overlap is reduced and redundant regions increase, we adjusted the hyperparameter ρ, which controls the overlap rate between input images, from 32 to 48. This setup allows us to evaluate AQFormer's performance under conditions with increased redundancy and reduced overlap.

The quantitative analysis results are shown in Table III. It is observed that AQFormer demonstrates more significant benefits in handling redundancy by selectively focusing on informative regions and suppressing less relevant ones. Furthermore, We notice that in hard scenarios, the variant "w/o GIM" outperforms the 1-scale AGNet. This suggests that iterating AQFormer only three times for large baselines is insufficient, as there still exist long-range dependencies between the two images that GIM cannot capture. Therefore, we propose a new variant "1-scale AGNet+" with M and N set to 6, 6. "1-scale AGNet+" demonstrates the best performance regardless of the scenario.

Figure 2 presents the visual comparison results. As the geometric transformation between input images increases, the variant without AQFormer (w/o AQFormer) exhibits noticeable artifacts. The variants 'w/o GIM' leveraging the global information provided by AQFormer, significantly reduce these artifacts. This demonstrates AQFormer's effectiveness in addressing substantial geometric changes. Additionally, the complete AGNet model with both AQFormer and GIM achieves superior performance, as it combines global and local information, further reducing artifacts and improving

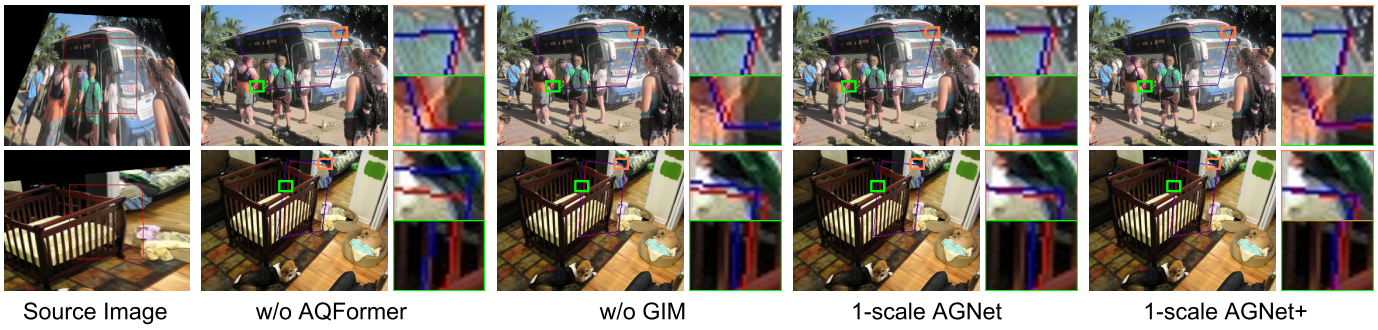


Fig. 14. Visualization of ablation results in the large baseline scenario.

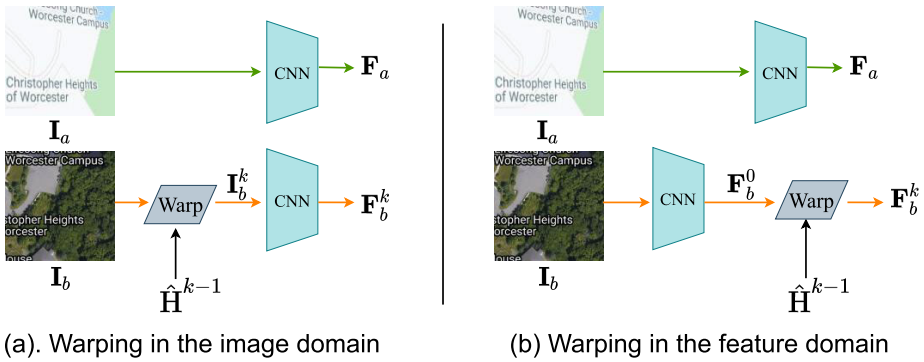


Fig. 15. Two methods for warping $\mathbf{F}_b$ to $\mathbf{F}_a$ under the guidance of $\hat{\mathrm{H}}^{k-1}$.

homography estimation under challenging conditions. Furthermore, "1-scale AGNet+" achieves the best performance, indicating that increasing the number of iterations is essential for improving visual quality in large baseline scenarios.

E. Study of Iterations

Iterative methods refine the homography matrix step by step, gradually reducing errors with each iteration. Intuitively, if the geometric transformation between the input image pair is small, only a few iterations are needed to estimate a reasonably accurate homography matrix. However, in large baseline scenarios, more iterations are required to refine the homography matrix gradually. To verify this hypothesis, we conducted experiments by adjusting the hyperparameter ρ and the number of iterations M and N. The larger the value of ρ, the lower the overlap between the input pairs.

The experimental results, as shown in Table IV, indicate that for small baseline scenarios (e.g., $\rho = 24$), only a few iterations are needed to achieve optimal performance. In contrast, additional iterations are beneficial for large scenarios (e.g., $\rho = 48$), as they enable the model to progressively refine the alignment between input images.

F. Study of Warping Methods

We observe that during the iterative process, $\mathbf{I}_b$ can be warped in two ways: warping in the image domain and warping in the feature domain, which is shown in Fig. 15. First, we introduce the detailed process of image domain warping: as shown in Fig. 15 (a), in the k-th iteration, $\mathbf{I}_b$ is warped into $\mathbf{I}_b^k$ guided by $\hat{\mathrm{H}}^{k-1}$, achieving a coarse alignment with $\mathbf{I}_a$. Then, $\mathbf{I}_a$ and $\mathbf{I}_b^k$ are fed into a convolutional network to obtain the features $\mathbf{F}_a$ and $\mathbf{F}_b^k$. Next, we introduce the detailed process of feature domain warping: as shown in Fig. 15 (b), $\mathbf{I}_a$ and $\mathbf{I}_b$ are fed into a convolutional network to obtain the features $\mathbf{F}_a$ and $\mathbf{F}_b^0$. $\mathbf{F}_b^0$ is warped into $\mathbf{F}_b^k$ guided by $\hat{\mathrm{H}}^{k-1}$ in the k-th iteration.

For feature domain warping, it is only necessary to extract features $\mathbf{F}_a$ and $\mathbf{F}_b^0$ during the first iteration. In contrast, features must be re-extracted after each warp for image domain warping. As a result, feature domain warping offers a lower computational cost. However, it is important to note that homography transformations involve translation, rotation, scaling, and other transformations, while convolutional neural networks lack rotation and scale invariance. Thus, applying homography transformation before feature extraction yields different results compared to performing feature extraction first, followed by the homography transformation. To further explore the impact of these two warping methods, we con-

TABLE IV

THE RELATIONSHIP BETWEEN OVERLAP RATE AND THE NUMBER OF ITERATIONS, WITH MACE ↓ AS THE EVALUATION METRIC. THE LARGER THE VALUE OF ρ, THE LOWER THE OVERLAP BETWEEN THE INPUT PAIRS. LARGER BASELINE SCENARIOS REQUIRE MORE ITERATIONS FOR REFINEMENT. M AND N REPRESENT THE NUMBER OF ITERATIONS OF AQFORMER AND GIM, RESPECTIVELY

	$M, N = 1, 1$	$M, N = 3, 3$	$M, N = 6, 6$
$\rho = 24$	0.3095	**0.0438**	0.0446
$\rho = 32$	0.5533	**0.0619**	0.0639
$\rho = 48$	6.9505	2.6772	**2.2097**

TABLE V
QUANTITATIVE COMPARISON OF MACE↓ ON THE TWO WARPING METHODS

Methods	Easy	Medium	Hard	Average
Warping in the feature domain	0.1123	0.1887	0.3566	0.2162
Warping in the image domain	**0.0230**	**0.0463**	**0.1215**	**0.0619**

ducted detailed ablation experiments. As shown in Table V, the experimental results indicate that warping in the feature domain leads to significant performance degradation. Therefore, we choose to warp $\mathbf{I}_b$ in the image domain.

VI. LIMITATIONS

While AGNet is effective across various scenarios, certain limitations need to be addressed. One specific challenge arises in cases where there are significant viewpoint transformations between image pairs. In such situations, the model may struggle to establish correspondences, leading to suboptimal performance. This difficulty can stem from the drastic changes in perspective, which can obscure common features and make it challenging for the model to accurately identify matching regions. To mitigate this issue, we are exploring various strategies that may improve the model's ability to adapt to varying perspectives and better capture the underlying geometric relationships between images.

VII. CONCLUSION

Long-range contexts are crucial for homography estimation. Recent approaches use transformers to capture long-range contexts. However, they ignore the phenomenon of query over-focusing. In this paper, we proposed an iterative network (AGNet) for homography estimation. AGNet predicts homographs coarse to fine in a global-to-local manner. To alleviate query over-focusing in capturing global contexts, we developed an adaptive query transformer that suppresses ambiguous queries and promotes key queries. In addition, we developed a gated interaction module to capture local contexts. Experimental results show that the proposed method performs favorably against state-of-the-art methods in general, cross-resolution, and cross-modal scenarios. In the future, we will explore more applications in tasks with fewer labeled data, such as estimating the homography of MRI and CT pairs.

REFERENCES

[1] S. Liu et al., "Content-aware unsupervised deep homography estimation and its extensions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 2849–2863, Mar. 2023.

[2] L. Nie, C. Lin, K. Liao, S. Liu, and Y. Zhao, "Parallax-tolerant unsupervised deep image stitching," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 7399–7408.

[3] X. Xiang, R. Abdein, N. Lv, and A. E. Saddik, "InvFlow: Involution and multi-scale interaction for unsupervised learning of optical flow," *Pattern Recognit.*, vol. 145, Jan. 2024, Art. no. 109918.

[4] J. Liu et al., "A dense light field reconstruction algorithm for four-dimensional optical flow constraint equation," *Pattern Recognit.*, vol. 134, Feb. 2023, Art. no. 109101.

[5] H. Xu, J. Liao, H. Liu, and Y. Sun, "Learning semantic alignment using global features and multi-scale confidence," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 2, pp. 897–910, Feb. 2024.

[6] Z. Song, C. Jia, L. Yang, H. Wei, and L. Liu, "GraphAlign++: An accurate feature alignment by graph matching for multi-modal 3D object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 2619–2632, Apr. 2024.

[7] X. Tang, Q. Yang, X. Zhang, W. Deng, H. Wang, and X. Gao, "A refinement method for single-stage object detection based on progressive decoupled task alignment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 5, pp. 3383–3394, May 2024.

[8] J. Ye, E. Pan, and W. Xu, "Digital video stabilization method based on periodic jitters of airborne vision of large flapping wing robots," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 2591–2603, Apr. 2024.

[9] W. Yan, Y. Sun, W. Zhou, Z. Liu, and R. Cong, "Deep video stabilization via robust homography estimation," *IEEE Signal Process. Lett.*, vol. 30, pp. 1602–1606, 2023.

[10] K. Matzen, M. F. Cohen, B. Evans, J. Kopf, and R. Szeliski, "Low-cost 360 stereo photography and video capture," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–12, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:13602603

[11] D. J. Brady et al., "Multiscale gigapixel photography," *Nature*, vol. 486, no. 7403, pp. 386–389, Jun. 2012.

[12] F. Yang, Y. Zhao, and X. Wang, "Common pole–polar properties of central catadioptric sphere and line images used for camera calibration," *Int. J. Comput. Vis.*, vol. 131, no. 1, pp. 121–133, 2023.

[13] S.-Y. Cao et al., "Recurrent homography estimation using homography-guided image warping and focus transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 9833–9842.

[14] R. Shao, G. Wu, Y. Zhou, Y. Fu, L. Fang, and Y. Liu, "LocalTrans: A multiscale local transformer network for cross-resolution homography estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14890–14899.

[15] S. Peleg, B. Rousso, A. Rav-Acha, and A. Zomet, "Mosaicing on adaptive manifolds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1144–1154, Oct. 2000.

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[17] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2548–2555.

[18] M. A. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[19] D. Barath, J. Matas, and J. Noskova, "MAGSAC: Marginalizing sample consensus," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10197–10205.

[20] Q. Jia et al., "Leveraging line-point consistence to preserve structures for wide parallax image stitching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 12186–12195.

[21] W. Xue, W. Xie, Y. Zhang, and S. Chen, "Stable linear structures and seam measurements for parallax image stitching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 253–261, Jan. 2022.

[22] M. Liu, S. Lin, H. Zhang, Z. Zha, and B. Wen, "Intrinsic-style distribution matching for arbitrary style transfer," *Knowl.-Based Syst.*, vol. 296, Jul. 2024, Art. no. 111898.

[23] G. Wu, X. Ning, L. Hou, F. He, H. Zhang, and A. Shankar, "Three-dimensional softmax mechanism guided bidirectional GRU networks for hyperspectral remote sensing image classification," *Signal Process.*, vol. 212, Nov. 2023, Art. no. 109151.

[24] W. Luo, H. Zhang, J. Li, and X.-S. Wei, "Learning semantically enhanced feature for fine-grained image classification," *IEEE Signal Process. Lett.*, vol. 27, pp. 1545–1549, 2020.

[25] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 224–236.

[26] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning feature matching with graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4938–4947.

[27] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "LoFTR: Detector-free local feature matching with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8922–8931.

[28] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Deep image homography estimation," 2016, *arXiv:1606.03798*.

[29] T. Nguyen, S. W. Chen, S. S. Shivakumar, C. J. Taylor, and V. Kumar, "Unsupervised deep homography: A fast and robust homography estimation model," *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 2346–2353, Jul. 2018.

[30] N. Ye, C. Wang, H. Fan, and S. Liu, "Motion basis learning for unsupervised deep homography estimation with subspace projection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13097–13105.

[31] L. Nie, C. Lin, K. Liao, S. Liu, and Y. Zhao, "Unsupervised deep image stitching: Reconstructing stitched features to images," *IEEE Trans. Image Process.*, vol. 30, pp. 6184–6197, 2021.

[32] X. Feng, Q. Jia, Z. Zhao, Y. Liu, X. Xue, and X. Fan, "Edge-aware correlation learning for unsupervised progressive homography estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 6, pp. 4773–4785, Jun. 2024.

[33] L. Nie, C. Lin, K. Liao, S. Liu, and Y. Zhao, "Depth-aware multi-grid deep Homography estimation with contextual correlation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4460–4472, Jul. 2021.

[34] J. Liu and X. Li, "Geometrized transformer for self-supervised homography estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 9556–9565.

[35] C.-H. Chang, C.-N. Chou, and E. Y. Chang, "CLKN: Cascaded Lucas–Kanade networks for image alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2213–2221.

[36] Y. Zhao, X. Huang, and Z. Zhang, "Deep Lucas–Kanade homography for multimodal image alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15950–15959.

[37] S.-Y. Cao, J. Hu, Z. Sheng, and H.-L. Shen, "Iterative deep homography estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1879–1888.

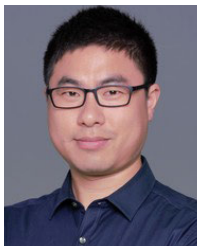
[38] M. Hong, Y. Lu, N. Ye, C. Lin, Q. Zhao, and S. Liu, "Unsupervised homography estimation with coplanarity-aware GAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17663–17672.

[39] H. Liu, Z. Dai, D. So, and Q. V. Le, "Pay attention to MLPs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 9204–9215.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[41] H. Zhu et al., "MCNet: Rethinking the core ingredients for accurate and efficient homography estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 25932–25941.

Zhongyang Li received the bachelor's degree in computer science from East China Normal University, Shanghai, China, in 2021, where he is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology. His research interests include image alignment and multimodal alignment.



Faming Fang received the Ph.D. degree in computer science from East China Normal University, Shanghai, China, in 2013. He is currently a Professor with the Department of Computer Science and Technology, East China Normal University. His research interests include image processing using the variational methods, PDEs, and deep learning methods.

Tingting Wang received the Ph.D. degree in computer science from East China Normal University, Shanghai, China, in 2021. She is currently a Post-Doctoral Fellow with the Department of Computer Science, East China Normal University. Her research interests include image processing using mathematical methods and deep learning methods.

Guixu Zhang received the Ph.D. degree from the Institute of Modern Physics, Chinese Academy of Sciences, Lanzhou, China, in 1998. He is currently a Professor with the Department of Computer Science and Technology, East China Normal University, Shanghai, China. His research interests include hyperspectral remote sensing, image processing, and artificial intelligence.