# Unsupervised Image Sequence Registration and Enhancement for Infrared Small Target Detection

Runze Hou[1], Puti Yan[2], Xuguang Duan[3], Xin Wang[3,*], *Member, IEEE*,

***Abstract*—In the burgeoning realm of deep learning and with the introduction of the infrared target detection dataset, infrared small target detection has increasingly garnered attention. Within this domain, multi-frame infrared small target detection stands as a pivotal and challenging sub-task. Notably, some recent methods have exhibited commendable performance in multi-frame infrared scenes. However, these methods were tethered to time-consuming background alignment pre-processing, which impedes their real-world application of multi-frame infrared target detection systems. In this paper, an unsupervised end-to-end framework tailored for infrared image sequence registration was proposed. This framework diverges from traditional registration methods by incorporating a novel Basket-based Hierarchical Temporal Consistency loss. The proposed loss function achieve intra-basket consistency and inter-basket diversity, effectively mitigating issues related to inconsistency. Additionally, the framework includes the Input Thresholding Mask and Output Transformation Mask. These components are crucial for guiding the network's training process and correcting misalignments. Moreover, the introduction of a dual-level residual enhancer is proposed to enhance the quality of registered images, thereby improving overall performance. Extensive experimental results have demonstrated the superiority of the proposed method over baseline methods. The proposed method achieved a significant improvement in the $F_1 - score$ metric on a public dataset, reaching 0.8882, and an inference speed of 23.34 FPS. This represents an improvement of 0.0190 in performance and a sixfold increase in speed compared to the state-of-the-art method in multi-frame infrared small target detection.**

***Index Terms*—Deep learning, computer vision, infrared imaging, target detection, image registration, image enhancement**

## I. Introduction

In recent years, the development of deep learning (DL) technologies has led to their integration into various research domains to enhance performance. Its integration is particularly evident in computer vision tasks, such as image classification [1], [2], object detection [3], semantic segmentation [4], and instance segmentation [5], [6]. Moreover, research interest has extended beyond natural images to encompass data from infrared, hyperspectral, and LiDAR sources due to their imaging advantages [7], [8]. As an important part of night rescue and ground observation, infrared small target detection also followed this trend, and the publication of several infrared datasets [9]–[13] for target detection has further accelerated this process. In the realm of infrared imaging, the detection of small targets presents specific challenges. Typically, these targets appear dim and are often obscured by complex and variable backgrounds, resulting in reduced texture and shape visibility and increased background noise. DL-based methods have shown effectiveness in addressing these challenges. They leverage extensive data to distinguish actual targets from background noise and maintain consistent performance across various background types. In contrast, the traditional methods required parameter tuning for each specific scenario, which limits their application in complex real-world systems.

Infrared small target methods can be categorized into two primary types: single-frame-based and multi-frame-based methods. The former focused on distinguishing targets from the background by local contrast information, while the latter required extracting spatio-temporal features through infrared sequences. Single-frame-based methods were suitable for the condition that there are significant distinctions in brightness, texture, and other factors between the targets and the background. However, the contrast information was insufficient when the targets become smaller and the background becomes more complex. In some cases, the motion information from multiple frames should be introduced to serve as the essential feature for detection. In traditional methods, researchers usually designed methods to exploit this information explicitly. For example, single-frame methods could utilize local contrast information through spatial filters, and multi-frame methods could utilize motion information through optical flow. In addition to these traditional methods, the research of single-frame infrared small target detection in DL area has made significant progress since the continuous efforts and exploration [9]–[11], [14]–[17] of researchers. Concurrently, the more complex scenario has garnered increasing interest, as evidenced by research focusing on multi-frame detection [7], cross-domain detection [18], and low-contrast scene detection [19].

In network design, the primary distinction between single-frame and multi-frame detectors lies in the requirement of multi-frame methods to capture motion information of moving targets across continuous frames. Though optical flow is an effective tool for motion modeling, its application is limited in small target detection due to the establishment condition of the motion consistency assumption cannot be satisfied by small targets in some scenarios. Therefore, the recent multi-frame networks [7], [20]–[22] directly took inter-frame difference after registration or alignment as the input motion information. The registration component in this process is critical. Without proper registration, the inter-frame difference will become the brightness change of the background rather than the targets.

1. Runze Hou is with Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, China. (hrz21@mails.tsinghua.edu.cn)

2. Puti Yan is with the Department of Aerospace Engineering, Harbin Institute of Technology, Harbin, China.

3. Xuguang Duan, Xin Wang is with the Department of Computer Science and Technology, Tsinghua University, China. (xin_wang@tsinghua.edu.cn)

* Corresponding author.

Therefore, the quality of registration profoundly influences the overall detection performance. It should be noted that owing to the distinctions in imaging techniques, natural images frequently exhibit higher resolutions and more details. Consequently, algorithms for the detection of small moving targets in natural images generally do not require a registration process. In contrast, datasets derived from infrared imaging, characterized by lower clarity and smaller size of targets, necessitate a rigorous reliance on the registration process for the effective detection of such small moving targets. In these multi-frame infrared small target detection methods, the standard registration pipeline comprises Feature Extraction, Feature Description, Feature Matching, and Homography Calculation. These stages are integral to various applications, including image registration and image stitching.

However, this pipeline works under a pair of images. This approach results in the division of continuous frames into several independent pairs for registration, consequently overlooking the continuity between adjacent frames. The challenges of this traditional pipeline can be categorized as follows:

- **Inconsistency**. The pairwise processing ignored the consistent motion of adjacent frames, and the outlier estimation will bring abnormal registration results, which result in wrong detection.
- **Misalignment**. The traditional pipeline relied on the extracted features from the original images. In instances where infrared images lack sufficient edge, texture, or other prominent visual information, the registration process was prone to failure.
- **High Time Consumption**. Each component of the pipeline required much computation, and combining the four parts will consume significant computing resources, making it difficult for the detection algorithms to process in real-time.

In this paper, an unsupervised image registration framework for infrared image sequences was proposed to address the identified weaknesses and limitations in existing registration pipelines. The main contributions of this paper can be summarized as follows:

- This paper proposed the first learning-based image sequence registration method for infrared images. The proposed method substantially enhanced the speed of sequence registration and solved the challenge of **High Time Consumption**. Together with the state-of-the-art (SOTA) detector, it offered a solution for a real-time infrared small moving target detection on edge AI devices.
- This paper introduced Basket-based Hierarchical Temporal Consistency constraint, which consisted of Intra-Basket Consistency and Inter-Basket Diversity strategy. The constraint addressed the issue of **Inconsistency** between adjacent frames and improved feature diversity.
- This paper designed a mask-guided mechanism, including Input Thresholding Mask and Output Transformation Mask for guiding both feature extraction and the network training. The mechanism solved the problem of **Misalignment** and addressed issues such as the lack of edges, as well as blurred texture encountered in infrared imaging.
- This paper proposed a Dual-level Residual Enhancer module to enhance the details in infrared images without multiplying the resolution of the input images.

## II. Related Works

### A. Homography Estimation

Homography estimation is the core component of image registration, where the image registration is the result by applying estimated homography transformation to the input image. Traditionally, this process comprises several stages: feature extraction (SIFT [23], SURF [24], and ORB [25]), feature description (SIFT [23], SURF [24], and ORB [25]), feature matching (FLANN [26]) and final estimation ( [27], RANSAC [28] and MAGSAC [29]). In recent years, DL-based alternatives for some of these steps have emerged, including SuperPoint [30] for feature extraction and description, and HardNet [31] for feature matching.

Apart from the traditional pipeline, various end-to-end methods have also been developed. DeTone et al. [32] proposed the first deep neural network HomographyNet for homography estimation, which achieves comparable performance with ORB. Nguyen et al. [33] first applied an unsupervised training style to deep homography estimation and achieved better performance compared with supervised methods under aerial images. Zhang et al. [34] further extended unsupervised homography estimation from aerial images to real-world images and proposed a new dataset for homography estimation. Ye et al. [35] proposed a new framework for unsupervised homography estimation, which estimates homography as a weighted sum of eight pre-defined orthogonal homography flow bases rather than four point offsets of a specific rectangle. Considering that a single homography can not represent the complex spatial transformation, Nie et al. [36] proposed a multi-grid deep homography network to predict homography from global to local. Hong et al. [37] additionally considered the problem of plane-induced parallax, proposed a generative adversarial network to add coplanar constraints and applyied a coarse-to-fine mechanism to predict homography transformation. Cao et al. [38] proposed an Iterative Homography Network (IHN), which predicts homography in an iterative manner under completely trainable parameters. Recently, Jiang et al. [39] proposed a progressive estimation strategy by converting large-baseline homography into multiple intermediate ones and training the whole network in a new semi-supervised style, which achieves better performance in large-baseline scenes.

There have also been methods focusing on image sequence registration. To mitigate error accumulation in inter-frame homography estimation, Liu et al. [40] utilized multiple reference frames for constructing panoramic images. Similarly, Lin et al. [41] and Kim et al. [42] employed high-resolution maps for error reduction. The above methods mainly solved the problem of error in panoramic image, not the temporal consistency of image registration. Dunau et al. [43] estimated the homography for infrared sequences through correspondences of GPS points from different frames. However, this method cannot utilize the original image to solve the problems, and was limited by the unavailability of GPS data in most scenarios.

Yang et al. [44] presented a novel integrated global-to-local framework that addresses the problems of dynamic infrared and visible image sequence registration. More recently, Zhao et al. [45] proposed an algorithm for registering planar infrared-visible image sequences through spatio-temporal association. Besides, there were several learning-based methods [46]–[50] that perform registration of infrared images. However, they ignored the temporal relation between continuous frames and required visible images for multi-modal fusion, which is not suitable for infrared small moving target detection. In contrast, the proposed method mainly solves the problem of deep homography estimation under infrared-only image sequences.

### *B. Infrared Small Target Detection*

Research on infrared small target detection has continued for decades. During this period, numerous researchers have proposed a wide array of methods grounded in diverse theoretical frameworks. This section aimed to review these methods for detecting infrared small targets, and introduce both traditional approaches and DL-based methods in recent years. Each category of methods is further divided into single-frame detection and multi-frame detection for discussion.

The main idea of the single-frame method was to use spatial contrast information to enhance the target and suppress the background noise. Current methods of utilizing spatial contrast information included imitating the human visual system [51], [52], designing spatial filters [53], performing saliency detection through attention mechanisms [54], locally modeling targets [55], and calculating gradient information [56].

In addition to spatial information, multi-frame methods also required to utilize spatiotemporal information, such as motion and brightness change, to deduce the location of the target. Current methods for extracting and utilizing spatiotemporal information included optical flow extraction [57], [58], background modeling [59], and spatiotemporal filtering [60].

Recently, some researchers have proposed new mechanisms for infrared small target detection. Cui et al. [61] designed a Hollow Side Window Filter (HSWF) to cope with the background estimation problem. This process was weighted by the saliency map constructed by heterogeneity filter. Zhou et al. [62] proposed a four-leaf model, which includes macro Background Suppressor (BS) and micro Texture Collector (TC). The combination of TC and BS not only suppress background clutter but also improve the detection performance of infrared small targets. Zhao et al. [63] first introduced the isolation Forest (iForest) mechanism into infrared small target detection area. By constructing both global iForest and local iForest, the proposed method improved the detection probability and eliminated the false alarms.

With the advent of deep learning technology and the introduction of infrared image datasets, learning-based neural networks have made great progress in infrared small target detection in recent years. These advancements have significantly reduced the need for scene-specific parameter adjustments inherent in traditional algorithms, significantly enhancing the applicability of infrared small target detection systems in varied real-world scenarios.

As mentioned above, infrared small target detection methods can be classified into single-frame and multi-frame methods. Among them, single-frame-based detection network design has been studied for a longer time.

In single-frame detection, the primary focus is on extracting local contrast information from an individual infrared frame. Most single-frame networks aim to extract and fuse contrast information at different scales for final prediction. For the multi-scale feature extraction part, some methods [15], [64]–[66] applied convolution kernels at different sizes to features on the same stage to achieve multi-scale feature extraction. In contrast, some methods [11] applied kernels at the same size to features from different stages to achieve multi-scale feature extraction. For the feature fusion part, there were progressive and parallel feature fusion mechanisms. The former [10], [15], [64] gradually up-sampled small-scale feature maps and fuses them with large-scale feature maps, while the latter [11], [65], [66] fused feature maps of all scales at once. Besides, some methods additionally designed different attention mechanisms for better extracting the features, such as Channel and Spatial Attention Module (CSAM) [11], Multi-head Self-Attention (MSA) [16]; Some methods designed different modulation mechanisms for better fusing the features, such as Asymmetric Contextual Modulation (ACM) [10], Bottom-up Local Attentional Modulation (BLAM) [15].

In addition, there were also several innovative solutions proposed to address detection challenges. Wang et al. [9] regarded infrared small target detection as a balance between miss detection and false alarm, and achieved the balance through adversarial training. Chen et al. [67] tried to use a unified framework to achieve both detection and segmentation tasks for utilizing location information better. Zhang et al. [19] designed a multiscale single-stage detector to handle scale changes of targets and proposed a nonlocal quadrature difference measure in deep feature space, which converting feature points that break semantic continuity to the potential target locations. Wang et al. [17] proposed Interior Attention-Aware network (IAANet). IAANet first obtained the region of interest through the Region Proposal Network (RPN) and then obtains prediction through attention perception. To reduce the gap between different domains, Zhang et al. [18] proposed an unsupervised domain adaptive method based on content decoupling, so as to better complete the detection of cross-domain small infrared targets. These various designs have significantly advanced single-frame infrared small target detection.

In contrast, there were still few detection networks specially designed for multi-frame scenes. But as the problems in single-frame scenes gradually solved, the problem of detection in multi-frame scenes has become a trend in recent years.

A straightforward solution for multi-frame detection was to input multiple frames into a single-frame detector. Some methods enhanced one frame with multi-frame information to use single-frame detectors. Kwan et al. [68] and Ying et al. [69] deployed super-resolution technology to enhance the target and fed the enhanced results to a single-frame detector for detection. However, these methods were not tailored for multi-frame scenarios, and the increasing resolution dramatically increases the computational consumption.

Some methods [70]–[72] modified single-frame detection networks to accept multiple frames by altering the input channels of first convolution layer. This kind of modification was direct but ignored the complex temporal relationship of continuous frames. After that, some methods [20]–[22] further replaced raw frames into frames after registration and fed the frame difference together with the raw frames to detection networks. These methods draw lessons from the frame difference method and have a stronger response to moving objects. Recently, Spatial-Temporal Differential Multi-scale Attention Network (STDMANet) [7] further introduced temporal multi-scale information to multi-frame detection methods.

As multi-frame infrared small target detection evolves, frame registration has become increasingly vital. However, as mentioned before, the traditional registration pipeline restricted the effectiveness and application potential of multi-frame infrared detectors. Although Li et al. [22] attempted to solve the problem of rapidity, this method was only designed to align two frames and not suitable enough for continuous sequences. Therefore, this paper aims to bridge these gaps.

## III. Proposed Method

This section outlines the structure of the proposed method. Section III-A is dedicated to presenting the preliminaries of deep homography estimation, including fundamental concepts and procedures. Subsequently, Sections III-B, III-C, and III-D each offer solutions designed to address the specific weaknesses inherent in the traditional registration pipeline. These sections collectively provide a comprehensive overview of the proposed framework and its underlying principles.

### A. Preliminary and Problem Formulation

For a deep homography estimation network, the input of the network are two gray images, source image $I_s$ and target image $I_t$ with the size $H \times W$. The expected results of deep homography estimation networks is homography flow/map $F_{st} \in \mathbb{R}^{H\times W\times 2}$, which represents the expected offset of each point in the image towards the horizontal and vertical direction from the source domain to the target domain. However, optimizing the problem of directly predicting the final homography is difficult [32], so most papers use intermediate variables to predict the final homography map.

Some methods [32]–[34] predict 4-point offsets $H_{4pt} \in \mathbb{R}^{4\times 2}$ of corner location as the intermediate variables, which is also called as 4-point parameterization. Given four points that can be connected as a quadrilateral in the source image, $H_{4pt}$ represents the offsets between the corresponding locations and original locations of these points. The offsets $H_{4pt}$ are then used to generate the non-singular homography matrix $H_{matrix} \in \mathbb{R}^{3\times 3}$ through the normalized Direct Linear Transform (DLT) algorithm [27]. Then, the homography matrix can be obtained from

$$F(x,y) = [H_{matrix} - I_{3\times 3}] \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{1}$$

where $x$ and $y$ denote the index of location in the horizontal and vertical direction, and the points are presented by homogeneous coordinates. Assuming that the prediction of the network is $H^*_{4pt}$, the supervised loss function used for backpropagation can be written as

$$L_H = \left\| H_{4pt} - H^*_{4pt} \right\|_p \tag{2}$$

where $p$ indicates the p-norm. However, since the ground truth $H_{4pt}$ is unavailable in most cases, the mainstream methods apply an unsupervised photometric loss for training. Assuming that $F^*_{st}$ is the estimated homography flow through DLT between $H^*_{4pt}$ and four original points, the unsupervised photometric loss from the source image to the target image can be written as

$$L^{st}_P = \left\| I_t - W(I_s, F^*_{st}) \right\|_p \tag{3}$$

where $W$ indicates warping $I_s$ to the perspective of $I_t$ according to the homography flow $F^*_{st}$.

Some methods [35], [37] predict weight factors for predefined orthogonal homography bases as the intermediate variables. The homography flow can be represented by eight orthogonal flow bases

$$F = \sum_i \alpha_i F_i, \qquad i = 1, 2, ..., 8 \tag{4}$$

where $F_i \in \mathbb{R}^{H\times W\times 2}$, $\alpha_i$ is coefficients and $\forall i, j, F_i^T F_j = 0$. For more details on bases acquisition, please refer to [35].

In addition to the photometric loss mentioned above, most methods [34], [35] apply feature loss $L_F$ at the same time. The feature of source image $I_s$ should be close to the feature of target image $I_t$ after registration. The loss $L^{st}_F$ can be formulated as

$$L^{st}_F = \left\| f_t - W(f_s, F^*_{st}) \right\|_p \tag{5}$$

where $f_s$ and $f_t$ is the extracted feature of $I_s$ and $I_t$, the superscript $s$ represents image from source domain, superscript $t$ represents image from target domain. Within the framework of the deep homography estimation network, the calculation of homography flow typically occurs simultaneously from the source domain to the target domain and vice versa during the forward process. Both of them are constrained by the loss function, that is, to calculate $L^{st}, L^{ts}$ at the same time and perform backpropagation. For the purpose of clarity in this exposition, the subsequent discussion will primarily focus on the perspective of transitioning from the source domain to the target domain.

The sequence registration problem for infrared small target detection has differences with above. It requires align $k-1$ frames $[I_{t-k+1}, ..., I_{t-1}]$ to the last frame $I_t$, where $k$ is the windows size and $t$ indicates time step. The expected results are $k-1$ homography flows $[F_{t-k+1\to t}, ..., F_{t-1\to t}]$. Then the aligned frames $\tilde{I_{t-i}}$ can be obtained through warping by homography flows, i.e.

$$\tilde{I_{t-i}} = W(I_{t-i}, F_{t-i\to t}), \qquad i = 1, 2, ..., k-1 \tag{6}$$

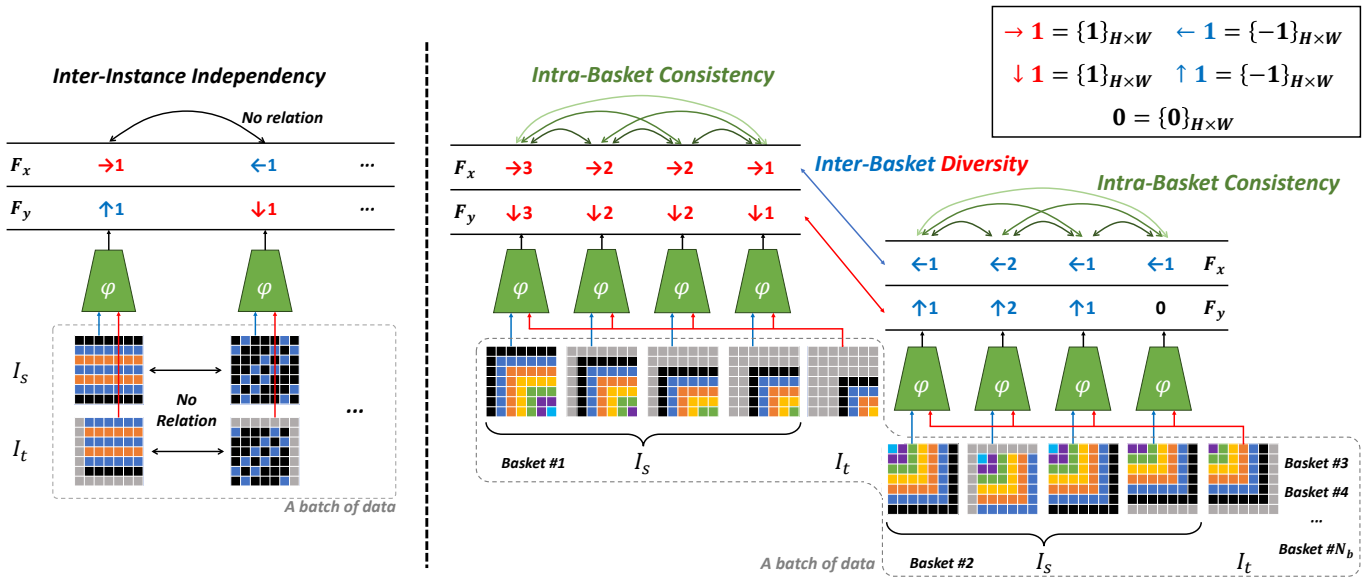


Fig. 1. The schematic diagram showing different strategies between other registration methods and the proposed method. The strategies of other registration methods can be summarized as Inter-Instance Independency strategy. In a batch of data, the strategy samples input data from the image pairs and estimates the homography flows between them. This strategy does not consider the cross-sample continuity or difference information and mines the negative sample from the single image pair. The proposed strategy are Intra-Basket Consistency and Inter-Basket Diversity strategy. This strategy divides a batch of input data into several baskets, each containing several consecutive video frames. By constraining the continuity of frames within the basket, consistency in estimating homography flow can be achieved. By expanding negative samples to different baskets, a diversity of negative samples can be achieved. The different textures of the input image represent infrared frames from different sequences. Among them, $I_s$ is the source domain image, $I_t$ is the target domain image, $\varphi$ is the homography flow estimation network, the output is the homography flow in $x$ direction and $y$ direction. The red output is the positive direction of homography flow, corresponding to the right direction of $x$ axis and downward direction of $y$ axis.

### *B. Basket-based Hierarchical Temporal Consistency*

Inspired by basket-based learning methods [73], a basket-based temporal consistency constraint is proposed to solve the problem of inconsistency. Because the input of multi-frame infrared detection is continuous, the consistency of background alignment is essential for the continuity of object motion, which is precisely the temporal information that multi-frame infrared detection pays the most attention to.

In the previous methods, the consideration of temporal consistency was not feasible since the input images are sampled from the image set rather than continuous videos. The homography flow is mainly estimated by distinguishing the transformed features and images from another view. This process typically accounted for symmetry by estimating both forward and backward homography flows within a given image pair, with each pair remaining unaffected by others. This strategy was referred to as Inter-Instance Independence in this paper. As shown in Fig. 1, this strategy only mines the positive and negative samples from a single image pair. Consequently, it failed to account for both temporal consistency and the relationships between various input samples.

In response to these limitations, a basket-based strategy was proposed. This approach involved dividing the input samples into several baskets, with each basket originating from a distinct infrared image sequence and containing a number of frames sampled from that sequence. The conventional methods of homography estimation equated to having a number of baskets equal to the number of sequences, where each basket comprised only two infrared frames (a source frame and a target frame). In the newly proposed strategy, as each basket encompassed consecutive infrared frames, it became possible to achieve temporal consistency in the output homography flow. This was accomplished by applying constraints to these consecutive frames, which is termed intra-basket consistency in this paper. On the other hand, if the input samples only contain frames from one infrared sequence, it is impossible to ensure that the model converges quickly and reasonably due to the high similarity of the input samples. To address this, and to further enhance the distinctiveness of homography flows, this paper extend negative samples from single image pairs to other baskets, which is referred as inter-basket diversity. The schematic diagram of our strategy is shown in Fig 1. To ensure both consistency within each basket and diversity among different baskets, it was necessary to set the number of baskets and the number of frames within each basket to relatively large values.

Specifically, the size of the input data batch $B_d$ is $N_b \times N_f \times H \times W$, where $N_b$ is the number of baskets and $N_f$ is the number of samples in each basket. In addition, since infrared frames can be viewed as grayscale images, there is no need to consider their colour channels. Afterwards, the first $N_f - 1$ frames in each basket are used as the source domain images, and the last frame is used as the target domain image for all previous frames, so the input size is changed to $N_b \times (N_f - 1) \times 2 \times H \times W$. After the homography estimation network, the output $B_f$ size is also $N_b \times (N_f - 1) \times 2 \times H \times W$. Note that in the third dimension, although there are two channels in both the input and output, the two channels of input represent the source and target domains, while the two channels of output represent the $x$ and $y$ directions of the homomorphy flow.

Based on the above definition, the basket-based hierarchical temporal consistency loss can be obtained from

$$L_{const.} = \frac{1}{N_b}\sum_{b=1}^{N_b}\sum_{Itv=1}^{N_f-2}\sum_{Idx=Itv}^{N_f-2} e^{-Itv}\|(B_f(b,Idx)-B_f(b,Idx-Itv)\|_p \tag{7}$$

where $Itv$ represents the interval that constrains continuity. An exponential decay strategy was utilized to demonstrate that the continuity constraint diminishes as the interval lengthens. $Idx$ is the index of frames within the basket, which was used for selecting two homography flows at predetermined intervals for the purpose of estimation.

In [35], triplet loss are introduced for feature contrast, which can be written as

$$\begin{aligned} L^{st}_{F,Trip.} =& \frac{1}{N_b*(N_f-1)}\sum_{b=1}^{N_b}\sum_{Idx=1}^{N_f-1}( \\ & \left\|B_f(b,N_f-1)-W(B_f(b,Idx),F^*_{Idx\to N_f-1})\right\|_p \\ & -\|B_f(b,N_f-1)-B_f(b,Idx)\|_p) \end{aligned} \tag{8}$$

The purpose of introducing triplet loss is to make the features after the transformation as close as possible to the features of the other perspective, while simultaneously maintaining dissimilarity in the original features of the two perspectives. Some contrastive learning methods [74], [75] further draw the conclusion that richness of negative sample is very important for representation learning. Consequently, the scope of negative samples was expanded from mere features of a single sample to the features of all other samples within the same batch, while preserving the fundamental purpose of the triplet loss. This augmented loss function was termed the Inter-Basket Diversity (IBD) loss

$$\begin{aligned} L^{st}_{F,IBDTrip.} =& \frac{1}{N_b*(N_f-1)}\sum_{b=1}^{N_b}\sum_{Idx=1}^{N_f-1}( \\ & \left\|B_f(b,N_f-1)-W(B_f(b,Idx),F^*_{Idx\to N_f-1})\right\|_p \\ & -\|B_f(b,N_f-1)-B_f(b,Idx)\|_p - L^{st}_{F,IBD} \end{aligned} \tag{9}$$

$$\begin{aligned} L^{st}_{F,IBD} =& \frac{1}{(N_b-1)*(N_f-1)}\sum_{b_d=1}^{N_b}\sum_{fIdx=1}^{N_f-1} \\ & [1-\delta(b,b_d)]\,\|B_f(b,N_f-1)-B_f(b_d,fIdx)\|_p \end{aligned} \tag{10}$$

where $\delta$ is an indicator function that has a value of 1 if $b$ equals $b_d$, used to exclude the current basket.

Ultimately, the loss function used for training is the sum of the following loss functions

$$\begin{aligned} L =& L^{st}_P + L^{ts}_P \\ & + \lambda_1(L^{st}_F + L^{ts}_F) \\ & + \lambda_2(L^{st}_{F,IBDTrip.} + L^{ts}_{F,IBDTrip.}) \\ & + \lambda_3(L^{st}_{const.} + L^{ts}_{const.}) \end{aligned} \tag{11}$$

where $\lambda_1, \lambda_2$ and $\lambda_3$ is the weights of corresponding losses.

### C. Mask as Guidance and Enhancement

In addition to inconsistency, another issue in traditional registration pipelines is misalignment due to the absence of distinct edge and texture information in some infrared images, complicating the feature extraction process for subsequent matching. To mitigate this issue, a mask-based mechanism was introduced. This mechanism involved thresholding the input to artificially generate precise edges and corners in blurry images. As a result, an adequate number of features were obtained after feeding both the original image and the additional mask into the network. This process was designated as Input Thresholding Mask $M_{IT}$, and its specific effect is illustrated in Figure 2. The process of feature extraction in this context can be described as follows.

$$f^s = \mathrm{FeatureProjection}([I_s, M^s_{IT}]) \tag{12}$$

Afterward, the features extracted from the original infrared frame and the generated thresholded mask were fed into the homography estimation network. Following the method outlined in [35], ResNet34 augmented with Low Rank Representation (LRR) blocks was utilized as the homography estimator. The output of the network comprised eight coefficients, serving as weights for predefined flow bases. By combining predefined homography flows with these weights, a homography flow from one image to another can be generated. Subsequently, the original image was resampled using this homography flow to achieve a perspective-transformed image. A crucial factor for the convergence of homography estimation networks was the comparison of these transformed images or features with the target images and their respective features. However, due to the varying degrees of background motion, in some cases, there may be large blank areas (typically 0) in the transformed images or features. It is unreasonable to calculate the loss in these areas. To address this, an Output Transformation Mask $M_{OT}$ was introduced.

$$M_{OT} = W(\mathbb{1}, F^*) \tag{13}$$

where $\mathbb{1}$ is all-ones matrix. The Output Transformation Mask is used to calculate loss $L_P, L_F$ and $L_{F,IBDTtip.}$ for detaching the loss of blank areas. The loss $L_P$ can be formulated as

$$L^{st}_P = M_{OT} \odot \|I_t - W(I_s, F^*_{st})\|_p\,, \tag{14}$$

and the loss $L_F$ and $L_{F,IBDTtip.}$ follow the same manner of guidance.

Previous research has shown that image enhancement, such as super-resolution, positively impact infrared target detection. However, the small size of infrared targets necessitates that the infrared target detection network maintains high-resolution feature maps of the original image size throughout the forward process of the network. The use of a super-resolution network significantly enlarges the size of the input image, which results in an exponential increase in the computational demands of the detection network. To address this issue, a local enhancement mechanism was introduced. This mechanism enhances the input image without altering the overall size of the images, thereby enhancing detection performance without escalating the computational load.

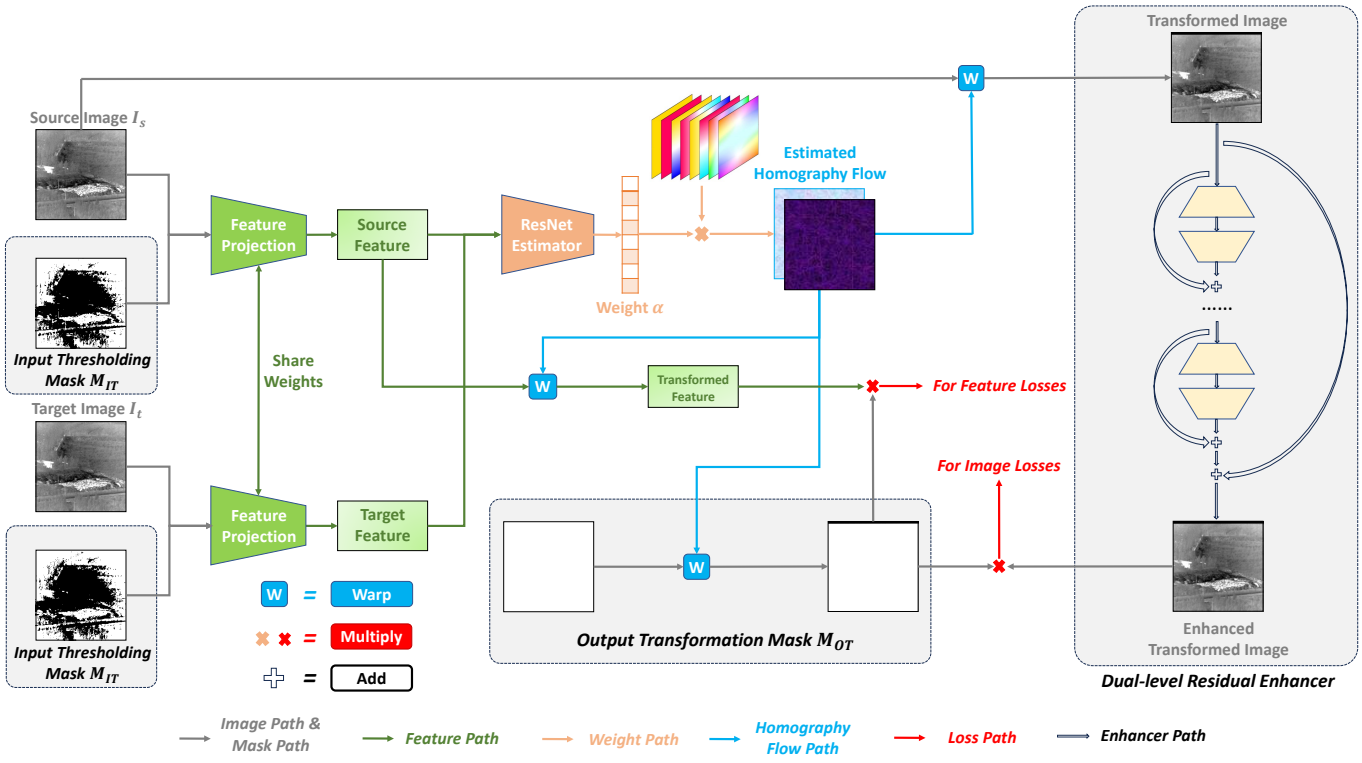


Fig. 2. Schematic diagram of the Mask-guided Local-enhancement Network (MgLeNet). The network first embeds the input frame and the Input Threshold Mask into the feature space and then obtains the predicted weights of predefined flow bases through the ResNet with LRR blocks to generate the homography flow. After transforming the frame through the homography flow, an additional two-level residual enhancer is used to further enhance the infrared frame.

In this paper, a dual-level residual module was designed to address this challenges. Firstly, learning the mapping from the original image to the enhanced image was difficult, typically necessitating a deep encoder-decoder network as opposed to a shallow structure. To overcome this, the first-level residual was introduced, enabling the network to focus solely on learning the updates of the image. Secondly, the updates to the image could be segmented into multiple smaller steps. To address this, the second-level residual was introduced, allowing the network to incrementally learn how to update images. The schematic diagram of this dual-level residual enhancer module is presented in Figure 2. More specifically, the dual-level enhancer comprised $e$ residual blocks, with the final enhancement output being the addition of the original image and the output from the final residual block. Each residual block was characterized as a combination of four ConvBlocks (CB)

$$f_{out} = f_{in} + \mathrm{CB}_1(\mathrm{CB}_8(\mathrm{CB}_{16}(\mathrm{CB}_8(f_{in})))) \qquad (15)$$

where each ConvBlock is Conv-BN-ReLU, and the subscript of CB is the output channel dimension of convolution layer.

The integration of mask guidance with local enhancement led to a homography flow estimation network, termed the Mask-guided Local-enhancement Network (MgLeNet). This network was instrumental in mitigating misalignment issues commonly encountered in infrared image registration and significantly enhanced the performance of infrared detection networks.

### D. Training Process and Details

The proposed method involved both the registration of infrared image sequences and the detection of infrared moving targets. Owing to the absence of suitable datasets and annotations for evaluating the registration performance, the performance of multi-frame infrared small target detection was utilized as an indirect measure to evaluate the efficacy of registration. Initially, the infrared homography flow estimation network was trained through the previously introduced loss function and network. Subsequently, this trained registration network was employed to fine-tune the detection network.

For the homography estimation network, the Adam [76] optimizer was applied to train the MgLeNet for 50 epochs with an initial learning rate of $10^{-5}$. An exponential decay method was used for adjusting the learning rate, with the decay rate $\gamma$ established at 0.8. The basket count $N_b$ was set to 8, and the sample frame count $N_f$ to 5, so training batch size was 32. Throughout the training, the original frames were resized to $320\times320$ and then cropped to $256\times256$ for data augmentation. In addition, the frames are randomly mirrored and normalized before being sent to the network. The parameter $\lambda_1, \lambda_2$ and $\lambda_3$ are set to 0.1, 0.1 and 0.001, respectively.

For the multi-frame detection network, STDMANet [7] offered the SOTA performance. The network are trained under SIFT+RANSAC registration pipelines. During the fine-tuning phase, the Adam optimizer was used to fine-tune the detection model with a batch size of 8 and an initial learning rate of $10^{-4}$ for 50 epochs.

## IV. EXPERIMENTS

In this section, extensive experiments were conducted to demonstrate the superiority of the proposed method. The dataset, metrics, and network details pertinent to these experiments were introduced in Sections IV-A and IV-B, respectively. In Section IV-C, the experimental results of the model were detailed and compared with baseline methods , encompassing both quantitative and qualitative aspects. Subsequently, the results of ablation experiments were presented in Section IV-D to illustrate the effectiveness of each module within the proposed method. Finally, the limitations of this paper are further discussed in Section IV-E.

### A. Dataset and Evaluation Metrics

In this paper, the DSAT [12] dataset was selected for the experimental evaluation. The DSAT dataset contains 22 sequences captured by infrared cameras from fixed-wing UAVs. Each sequence contains one or two targets, and the backgrounds of the dataset include sky and ground. The DSAT dataset consists of a total of 16,177 frames and 16,944 targets.

For the evaluation of the proposed method, the same metrics as those employed by STDMANet were followed. Infrared small targets were represented by their centers, and the distance between the predicted and ground truth centers was used as a criterion to determine the success of target detection. Under this definition, $Precision, Recall$, and $F_1 - score$ are calculated by the following equation

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

$$Recall = \frac{TP}{TP + FN} \tag{17}$$

$$F_1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{18}$$

It was stipulated that a prediction was considered correct ($TP$) if the distance between the predicted center and the ground truth center was within 10 pixels. Additionally, missed detections ($FN$) and false detections ($FP$) were added to calculate the corresponding metrics.

### B. Network Details and Settings

The MgLeNet employed ResNet34 with LRR blocks as its backbone network for predicting eight orthogonal flow bases. The calculation and acquisition of these bases can be referred to [35]. Additionally, the Feature Projection module of MgLeNet is composed of three convolution blocks, with output channels of 4, 8, and 1, respectively. The dual-level residual enhancer is composed of three residual blocks, and each residual block is composed of four convolution layers with channels of 8, 16, 8 and 1 respectively.

The STDMANet utilized for detection adheres to the structure outlined in the original paper. The length of the infrared frame sequence is configured to 20 frames, with each image sized at $256 \times 256$. The feature channels for the differential path, dynamic path, and static path are set to 96, 32, and 32, respectively, while the output channels for feature aggregation and Spatial Multi-Scale Feature Refiner are configured to 32.

Furthermore, various network settings were provided in the ablation studies. Initially, the basket-based hierarchical consistency constraint was omitted, resulting in two distinct settings: one without the consistency loss and another without the IBD loss. Subsequently, the mask guidance was removed, leading to two additional settings lacking the Input Thresholding Mask and the Output Transformation Mask. Lastly, a setting that excludes the dual-level residual enhancer was also examined.

All networks and settings are built by the Pytorch [77] framework, and are trained and tested on an Nvidia GeForce RTX 3090 GPU, supported by a 72-core Intel Xeon CPU.

### C. Comparison to Existing Methods

The proposed method was compared with both traditional and recent registration pipelines, focusing on the performance in multi-frame infrared small target detection. Two distinct comparison approaches were employed. The first approach directly tested the homography estimator on a detection network trained on the SIFT+RANSAC pipeline. The second approach fine-tuned the detection network with the new registration pipeline. The former was designed to assess the generalization capability of the registered pipeline to the original pipeline, whereas the latter aimed to evaluate the maximum potential performance of the registered pipeline. For the Infrared Homography Network (IHN), two different configurations were devised. Config 1 operated on a single scale (resolution) with six iterations, while Config 2 introduced a second scale and conducted three additional iterations at this scale.

**Quantitative comparison.** Table I shows the performance of the proposed approach compared with the baseline methods, where the proposed method shows the best performance. Compared with the state-of-the-art method STDMANet in multi-frame infrared small target detection, the proposed method can increase the metric $F_1$ score by 0.0190. In addition, the proposed method also has better performance than other registration algorithms. In the results without fine-tuning the detection network, IHN achieves the best performance, while the proposed method achieves sub-optimal results. This demonstrates that the performance of IHN registration is the most similar to that of the SIFT+RANSAC pipeline, while the proposed method is the second. Since the original model is trained with the SIFT+RANSAC registration pipeline, other pipelines can not perform better than the original pipeline. After fine-tuning the detection network, the proposed method performs best, and the second best is LoFTR. The highest detection performance indirectly shows that the proposed method is the best for the correctness of registration.

The performance comparison before and after fine-tuning on the SIFT+RANSAC pipeline revealed minimal differences. This was attributed to the fact that the original network had been trained on this pipeline, so the fine-tuning acted as an additional learning rate decay. However, for other registration pipelines, the detection performance showed varying degrees of improvement after fine-tuning. Notably, the $F_1$ scores of the two ORB-based pipelines did not exceed 0.8. Given that the ORB compromises feature quality for speed compared with the SIFT, this decline in quality had a significantly negative

TABLE I
**Target detection results under different registration pipelines.** THE HIGHEST $F_1$ SCORE IS IN BOLDFACE, AND THE SECOND HIGHEST IS UNDERLINED.

| Fine-tuning Registration | Running Registration | Precision ↑ | Recall ↑ | $F_1 - score$ ↑ |
|---|---|---|---|---|
| ✗ | ✗ | 0.7437 | 0.7685 | 0.7559 |
| ✗ | SIFT [23] + RANSAC [28] | 0.8378 | **0.9031** | 0.8692 |
| ✗ | SIFT [23] + MAGSAC [29] | 0.8316 | 0.8985 | 0.8637 |
| ✗ | ORB [25] + RANSAC [28] | 0.7084 | 0.8096 | 0.7556 |
| ✗ | ORB [25] + MAGSAC [29] | 0.7407 | 0.8335 | 0.7844 |
| ✗ | LoFTR [78] + RANSAC [28] | 0.8182 | 0.8832 | 0.8495 |
| ✗ | LoFTR [78] + MAGSAC [29] | 0.8176 | 0.8827 | 0.8489 |
| ✗ | IHN [38] Config 1 | 0.8040 | 0.8743 | 0.8377 |
| ✗ | IHN [38] Config 2 | 0.8281 | 0.8947 | 0.8601 |
| ✗ | Ours | 0.8232 | 0.8904 | 0.8555 |
| SIFT [23] + RANSAC [28] | SIFT [23] + RANSAC [28] | 0.8687 | 0.8772 | 0.8729 |
| SIFT [23] + MAGSAC [29] | SIFT [23] + MAGSAC [29] | 0.8700 | 0.8789 | 0.8744 |
| ORB [25] + RANSAC [28] | ORB [25] + RANSAC [28] | 0.7506 | 0.7863 | 0.7680 |
| ORB [25] + MAGSAC [29] | ORB [25] + MAGSAC [29] | 0.7822 | 0.8185 | 0.7999 |
| LoFTR [78] + RANSAC [28] | LoFTR [78] + RANSAC [28] | <u>0.8707</u> | 0.8801 | <u>0.8754</u> |
| LoFTR [78] + MAGSAC [29] | LoFTR [78] + MAGSAC [29] | 0.8698 | 0.8795 | 0.8746 |
| IHN [38] Config 1 | IHN [38] Config 1 | 0.8333 | 0.8737 | 0.8530 |
| IHN [38] Config 2 | IHN [38] Config 2 | 0.8514 | 0.8861 | 0.8684 |
| Ours | Ours | **0.8746** | <u>0.9022</u> | **0.8882** |

TABLE II
**Inference time and FPS between the proposed model and other methods.** THE TIME SHOWN HERE IS THE INFERENCE TIME OF THE WHOLE PROCESS (REGISTRATION AND DETECTION), NOT JUST THE INFERENCE TIME OF THE DETECTION NETWORK. HENCE, THE SPEED OF THE STDMANET, WHICH DEPENDS ON HEAVY CPU COMPUTATION DURING REGISTRATION, IS MUCH SLOWER THAN THE ORIGINAL PAPER.

| Method | Inference Time ↓ | FPS ↑ | Speed Up |
|---|---|---|---|
| SIFT [23] + RANSAC [28] | 259ms | 3.86 | 1× |
| SIFT [23] + MAGSAC [29] | 262ms | 3.82 | ∼ 1× |
| ORB [25] + RANSAC [28] | 89ms | 11.29 | 2.92× |
| ORB [25] + MAGSAC [29] | 89ms | 11.18 | 2.90× |
| LoFTR [78] + RANSAC [28] | 332ms | 3.01 | 0.78× |
| LoFTR [78] + MAGSAC [29] | 332ms | 3.00 | 0.78× |
| IHN [38] Config 1 | 82ms | 12.20 | 3.16× |
| IHN [38] Config 2 | 117ms | 8.58 | 2.22× |
| Ours | **43ms** | **23.34** | 6.05× |

impact on the detection of infrared small targets. After fine-tuning, the pipeline proposed in this paper exhibited the most substantial progress. Considering that the proposed pipeline was specifically designed to address the issue of temporal consistency in continuous registration, the observed performance gap indicated the importance of temporal consistency. This omission likely limited the potential for further enhancements in detection performance of other registration pipelines.

At the same time, the speed (FPS) and inference time of the proposed method compared to other methods is shown in Table II, where the batch size is set to 1 in all speed tests. Compared with the original STDMANet pipeline, the proposed method can increase the speed by more than 6 times and is also faster than the registration method used for comparison.

**Qualitative comparison.** In Fig.3, a sample sequence was displayed for the purpose of visualizing registration, along with the results of various registration methods applied to this sequence. For simplicity, only the results of MAGSAC were provided in this illustration since there is no essential difference between RANSAC and MAGSAC.

In most sequences, each registration pipeline can perform well. The sample sequence is a challenging scene that contains fog, which causes the features of the image to become blurred. The background motion in the sequence is mainly horizontal, with only a tiny vertical change. At the same time, the motion is mainly caused by translation and does not include more complex rotation, scaling, affine, and other transformations.

The original SIFT-based registration pipeline faces significant challenges from the registration results, especially in frame #13, which produces a significant tilt. At the same time, the registration result is inconsistent, and there is an apparent sudden change in the registration result of frame #15. As a faster alternative to SIFT, ORB performs worse, produces an incredible registration result at frame #14, and hardly changes in other frames since it does not detect enough feature points. LoFTR and IHN perform well, but there is a tilt simultaneously, while there are only translation transformations in the sequence. In contrast, the proposed method performs best, the transformations are all translations, and the consistency is also well-guaranteed. Furthermore, we have enlarged the registration result of frame #16 in the figure. For enlarged frames, the SIFT pipeline demonstrated significant

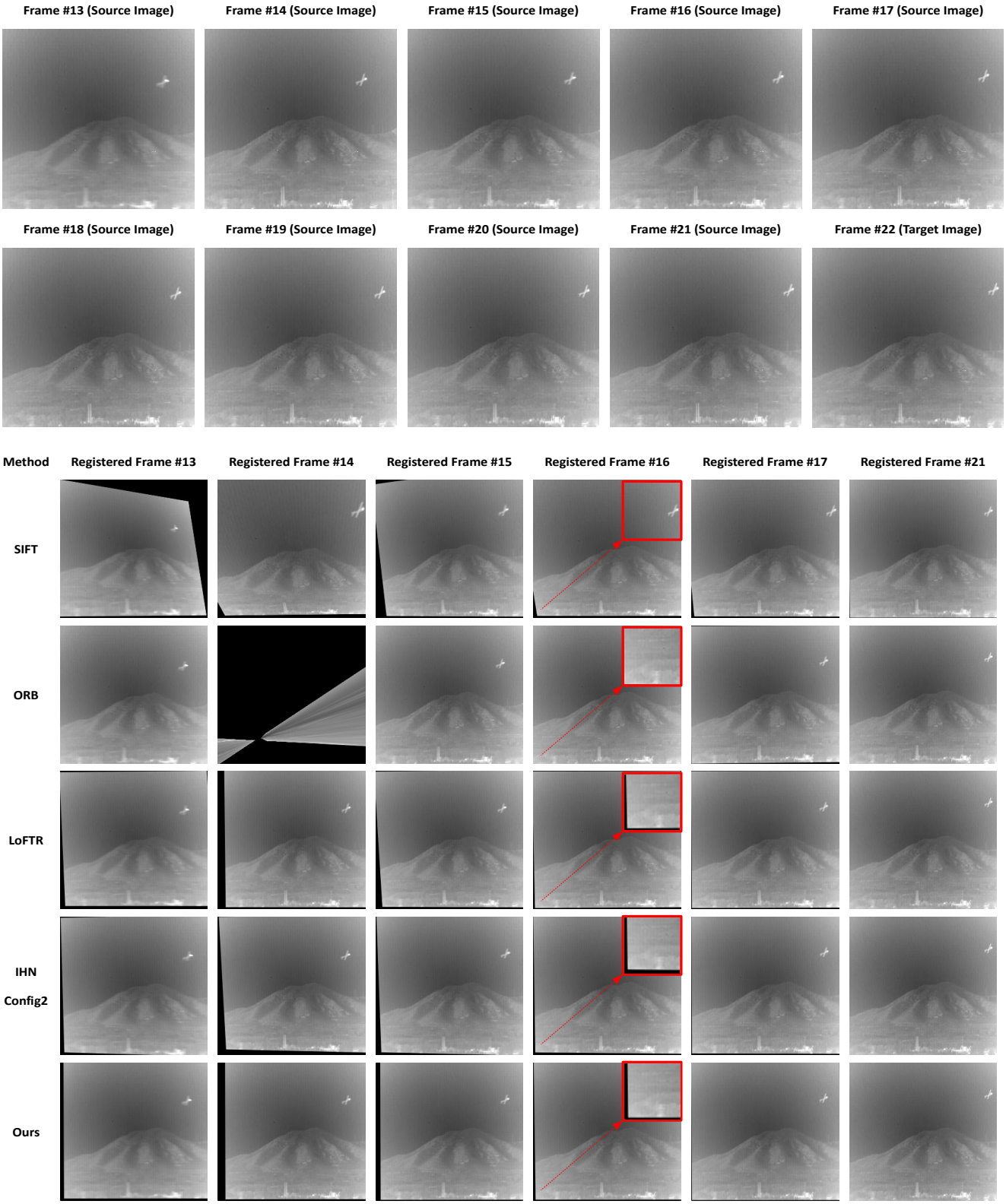


Fig. 3. Visualization results of different registered pipelines on an image sequence. In the top two rows, the original infrared image sequence was displayed, comprising a total of 10 frames. Among these, the first 9 frames represented images from the source domain, while the final frame served as the target for all source domain images. The comparison between the results of the proposed method and other pipelines was illustrated in the subsequent five rows. Given that the motion in the last few frames of the sequence was small, only the registration results of the first five frames and the penultimate frame were shown.

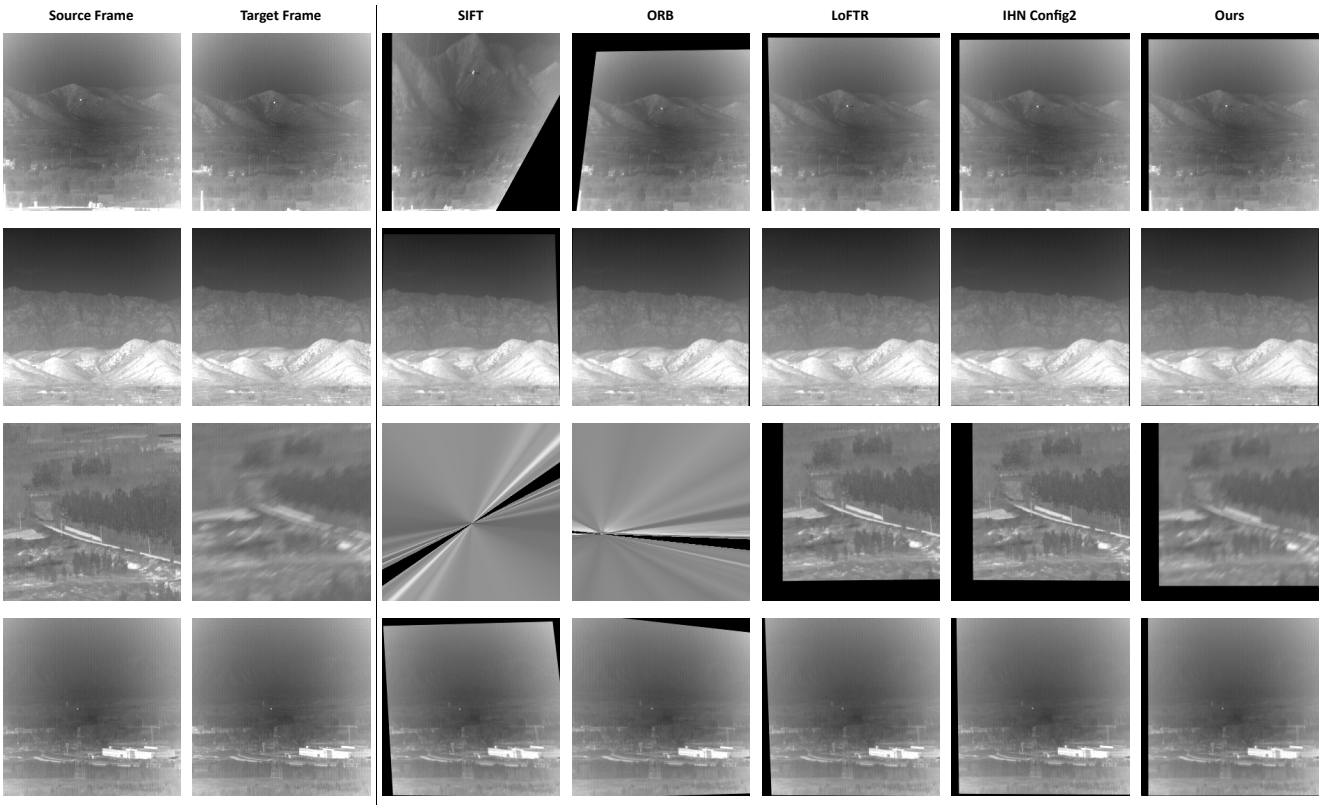


Fig. 4. Visualization results of different registered pipelines on several image pairs. The first two columns represented the source and target images. The comparison between the results of the proposed registration pipelines and other pipelines was illustrated in the subsequent five columns.

skew, the ORB pipeline failed to achieve registration, and the discrepancies for other pipelines were within few pixels. However, considering that the target size may be within few pixels for multi-frame infrared small target detection, pixel-level improvements have a positive effect on detection.

In addition to the above sequence, we also selected other challenging image pairs for presentation. These scenes included cities and mountains, and include conditions such as fog and motion blur. The registration results of different registration pipelines were shown in Fig. 4 . Comparing different registration pipelines, it was difficult for pipelines based on SIFT and ORB to achieve reasonable results in these scenarios. The LoFTR pipeline exhibited improved performance, but tended to produce skewed predictions for translations (first and fourth rows). IHN performed well in most cases, and was only slightly inferior to our method in few pixels. For the registration results in the second row, only our method shows a slight and stable upward movement trend of background.

Through Fig. 3 and Fig. 4, we qualitatively demonstrated the superiority of our method in terms of consistency and diversity respectively. Compared with existing registration pipelines, our proposed method could produce temporal consistent predictions and perform well in various complex scenarios.

### D. Ablation Studies

Several ablation studies were conducted on the DSAT dataset to ascertain the effectiveness of the components within the proposed method. The results of these experiments are presented in Table III. Specific modules were removed from the complete model, following which the detection performance was evaluated to determine their contributions to the model. The first row of the table displays the performance of the complete model, while the subsequent rows detail the ablation of various losses, masks, and modules. As indicated by the table, the removal of each component from the model resulted in a degradation of performance, thereby affirming the efficacy of the proposed module. In all ablation experiments, the removal of consistency loss has the greatest impact on the model, which further illustrates the importance of solving the problem of temporal consistency in multi-frame registration.

### E. Limitations

Some additional special cases were shown in Fig 5. The first two rows of the figure showed two examples of misalignment. The SIFT-based pipeline failed in both examples since the background was sky with few features. In contrast, the proposed method guided the network by input thresholding masks to generate additional salient features, which achieved better performance in the first example. In the second example, even adding additional masks could not guarantee enough feature information since the thresholding operation could not generate relatively stable masks. Hence, the proposed method and the SIFT-based pipeline all failed. However, this situation was also reasonable to some extent since it was almost impossible

TABLE III
**Ablation results on DSAT dataset.** THE PERFORMANCE GAP UNDER DIFFERENT ABLATION ON LOSSES, MASKS AND MODULES ARE SHOWN.

| $L_{const.}$ | $L_{F,IBDTrip.}$ | Input Thresholding Mask | Output Transformation Mask | Local Enhancement | Precision | Recall | $F_1$ score | $F_1$ Diff |
|---|---|---|---|---|---|---|---|---|
| ✓ | ✓ | ✓ | ✓ | ✓ | 0.8746 | 0.9022 | 0.8882 | 0.0000 |
| | ✓ | ✓ | ✓ | ✓ | 0.8482 | 0.8852 | 0.8663 | -0.0219 |
| ✓ | | ✓ | ✓ | ✓ | 0.8672 | 0.8947 | 0.8807 | -0.0075 |
| ✓ | ✓ | | ✓ | ✓ | 0.8678 | 0.8002 | 0.8837 | -0.0045 |
| ✓ | ✓ | ✓ | | ✓ | 0.8691 | 0.9019 | 0.8852 | -0.0030 |
| ✓ | ✓ | ✓ | ✓ | | 0.8553 | 0.8832 | 0.8690 | -0.0192 |

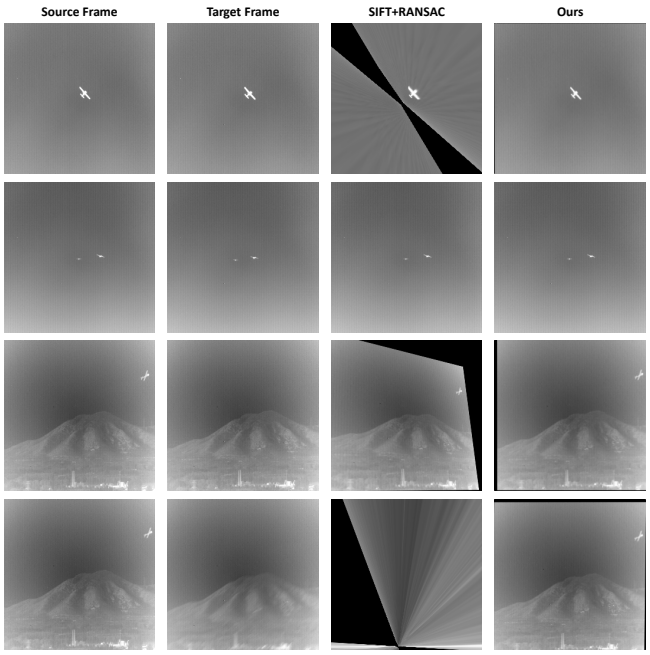


Fig. 5. Some special case examples for registration. Some registration results based on the proposed method and SIFT-based pipelines are presented to reveal the reasons and mechanisms of some registration failures.

for this scenario to achieve successful registration based on the registration theory. The last two rows in the figure show two examples of consistency, which have very similar source and target domain images. However, the registration results of the two examples were obviously different for the SIFT-based pipeline, further illustrating the importance of introducing consistency. This kind of registration results would seriously misunderstand the model and lead to wrong detection results. With consistency considered, the proposed method generated more stable results.

Consequently, it was observed that both traditional pipelines and DL-based registration methods were prone to failure in scenarios where the image contained insufficient features. The Input Thresholding Mask proposed in this paper is more suitable for scenes with complex but blurry background. Furthermore, the porposed registration method necessitated training on specific datasets. If the background was less blurry and time consumption was not the primary concern, tradtional registration pipelines might represent a more suitable choice.

To sum up, the proposed method solved the problem of inconsistency and misalignments in infrared image sequence registration. It could significantly accelerate the speed of multi-frame infrared target detection while slightly improving the detection performance. The proposed method have exceeded the recent registration pipelines both in terms of testing accuracy and running speed.

## V. CONCLUSION

Starting from the time-consuming background alignment process in multi-frame infrared small target detection, this paper summarizes three problems: inconsistency, misalignment, and high time consumption. In order to solve these problems, this paper designs a basket-based hierarchical consistency constraint and mask guidance to solve the problem of inconsistency and misalignment. In addition, an additional two-level residual enhancer is introduced to improve the detection performance further. Finally, this paper achieves a significant performance improvement and more than 6 times speed improvement compared with the original pipeline. It takes a big step toward the practical application of multi-frame infrared small target detection systems. However, using the end-to-end method to generate the registered image is still indirect. In the future, integrating the registration and detection networks to achieve an end-to-end detection process can further solve the problem of balance between accuracy and efficiency.

## REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[2] J. Wang, W. Li, M. Zhang, R. Tao, and J. Chanussot, "Remote sensing scene classification via multi-stage self-guided separation network," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.
[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.
[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.
[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.
[6] Y. Han, J. Zhang, Z. Xue, C. Xu, X. Shen, Y. Wang, C. Wang, Y. Liu, and X. Li, "Reference twice: A simple and unified baseline for few-shot instance segmentation," *arXiv preprint arXiv:2301.01156*, 2023.
[7] P. Yan, R. Hou, X. Duan, C. Yue, X. Wang, and X. Cao, "Stdmanet: Spatio-temporal differential multiscale attention network for small moving infrared target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[8] M. Zhang, W. Li, Y. Zhang, R. Tao, and Q. Du, "Hyperspectral and lidar data classification based on structural optimization transmission," *IEEE Transactions on Cybernetics*, 2022.

[9] H. Wang, L. Zhou, and L. Wang, "Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8509–8518.

[10] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Asymmetric contextual modulation for infrared small target detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 950–959.

[11] B. Li, C. Xiao, L. Wang, Y. Wang, Z. Lin, M. Li, W. An, and Y. Guo, "Dense nested attention network for infrared small target detection," *IEEE Transactions on Image Processing*, 2022.

[12] B. Hui, Z. Song, H. Fan, P. Zhong, W. Hu, X. Zhang, J. Ling, H. Su, W. Jin, Y. Zhang *et al.*, "A dataset for infrared image dim-small aircraft target detection and tracking under ground/air background," *Sci. Data Bank*, vol. 5, p. 12, 2019, https://www.scidb.cn/en/detail?dataSetId=720626420933459968&dataSetType=journal.

[13] X. Sun, L. Guo, W. Zhang, Z. Wang, Y. Hou, Z. Li, and X. Teng, "A dataset for small infrared moving target detection under clutter background," Feb. 2022, https://www.scidb.cn/en/detail?dataSetId=808025946870251520&dataSetType=journal.

[14] M. Zhao, L. Cheng, X. Yang, P. Feng, L. Liu, and N. Wu, "Tbc-net: A real-time detector for infrared small target detection using semantic constraint," *arXiv preprint arXiv:2001.05852*, 2019.

[15] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Attentional local contrast networks for infrared small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 11, pp. 9813–9824, 2021.

[16] F. Liu, C. Gao, F. Chen, D. Meng, W. Zuo, and X. Gao, "Infrared small-dim target detection with transformer under complex backgrounds," *arXiv preprint arXiv:2109.14379*, 2021.

[17] K. Wang, S. Du, C. Liu, and Z. Cao, "Interior attention-aware network for infrared small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[18] Y. Zhang, Y. Zhang, Z. Shi, R. Fu, D. Liu, Y. Zhang, and J. Du, "Enhanced cross-domain dim and small infrared target detection via content-decoupled feature alignment," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[19] Y. Zhang, Y. Zhang, R. Fu, Z. Shi, J. Zhang, D. Liu, and J. Du, "Learning nonlocal quadrature contrast for detection and recognition of infrared rotary-wing uav targets in complex background," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–19, 2022.

[20] J. Du, D. Li, Y. Deng, L. Zhang, H. Lu, M. Hu, X. Shen, Z. Liu, and X. Ji, "Multiple frames based infrared small target detection method using cnn," in *2021 4th International Conference on Algorithms, Computing and Artificial Intelligence*, 2021, pp. 1–6.

[21] J. Du, H. Lu, L. Zhang, M. Hu, S. Chen, Y. Deng, X. Shen, and Y. Zhang, "A spatial-temporal feature-based detection framework for infrared dim small target," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2021.

[22] D. Li, B. Mo, and J. Zhou, "Boost infrared moving aircraft detection performance by using fast homography estimation and dual input object detection network," *Infrared Physics & Technology*, vol. 123, p. 104182, 2022.

[23] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91–110, 2004.

[24] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Lecture notes in computer science*, vol. 3951, pp. 404–417, 2006.

[25] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*. Ieee, 2011, pp. 2564–2571.

[26] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration." *VISAPP (1)*, vol. 2, no. 331-340, p. 2, 2009.

[27] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

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

[29] D. Barath, J. Matas, and J. Noskova, "Magsac: marginalizing sample consensus," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 197–10 205.

[30] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 224–236.

[31] A. Mishchuk, D. Mishkin, F. Radenovic, and J. Matas, "Working hard to know your neighbor's margins: Local descriptor learning loss," *Advances in neural information processing systems*, vol. 30, 2017.

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

[33] 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 Robotics and Automation Letters*, vol. 3, no. 3, pp. 2346–2353, 2018.

[34] J. Zhang, C. Wang, S. Liu, L. Jia, N. Ye, J. Wang, J. Zhou, and J. Sun, "Content-aware unsupervised deep homography estimation," in *European Conference on Computer Vision*. Springer, 2020, pp. 653–669.

[35] N. Ye, C. Wang, H. Fan, and S. Liu, "Motion basis learning for unsupervised deep homography estimation with subspace projection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 13 117–13 125.

[36] L. Nie, C. Lin, K. Liao, S. Liu, and Y. Zhao, "Depth-aware multi-grid deep homography estimation with contextual correlation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4460–4472, 2021.

[37] M. Hong, Y. Lu, N. Ye, C. Lin, Q. Zhao, and S. Liu, "Unsupervised homography estimation with coplanarity-aware gan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 663–17 672.

[38] S.-Y. Cao, J. Hu, Z. Sheng, and H.-L. Shen, "Iterative deep homography estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1879–1888.

[39] H. Jiang, H. Li, Y. Lu, S. Han, and S. Liu, "Semi-supervised deep large-baseline homography estimation with progressive equivalence constraint," *arXiv preprint arXiv:2212.02763*, 2022.

[40] G. Liu, X. Tang, J. Huang, J. Liu, and D. Sun, "Hierarchical model-based human motion tracking via unscented kalman filter," in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[41] Y. Lin and G. Medioni, "Map-enhanced uav image sequence registration and synchronization of multiple image sequences," in *2007 IEEE conference on computer vision and pattern recognition*. IEEE, 2007, pp. 1–7.

[42] H. Kim and S. Han, "Geo-registration of wide-baseline panoramic image sequences using a digital map reference," *Multimedia Tools and Applications*, vol. 76, pp. 11 215–11 233, 2017.

[43] P. Dunau, D. Fitz, and J. Beyerer, "Homography estimation for low-contrast ir image sequences utilizing gps control points," *tm-Technisches Messen*, vol. 82, no. 5, pp. 262–272, 2015.

[44] M. Y. Yang, Y. Qiang, and B. Rosenhahn, "A global-to-local framework for infrared and visible image sequence registration," in *2015 IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2015, pp. 381–388.

[45] B. Zhao, T. Xu, Y. Chen, T. Li, and X. Sun, "Automatic and robust infrared-visible image sequence registration via spatio-temporal association," *Sensors*, vol. 19, no. 5, p. 997, 2019.

[46] M. Arar, Y. Ginger, D. Danon, A. H. Bermano, and D. Cohen-Or, "Unsupervised multi-modal image registration via geometry preserving image-to-image translation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 410–13 419.

[47] B. Debaque, H. Perreault, J.-P. Mercier, M.-A. Drouin, R. David, B. Chatelais, N. Duclos-Hindie, and S. Roy, "Thermal and visible image registration using deep homography," in *2022 25th International Conference on Information Fusion (FUSION)*. IEEE, 2022, pp. 1–8.

[48] T. Pouplin, H. Perreault, B. Debaque, M. Drouin, N. Duclos-Hindie, and S. Roy, "Multimodal deep homography estimation using a domain adaptation generative adversarial network," in *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, 2022, pp. 3635–3641.

[49] D. Wang, J. Liu, X. Fan, and R. Liu, "Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration," *arXiv preprint arXiv:2205.11876*, 2022.

[50] Z. Jiang, Z. Zhang, J. Liu, X. Fan, and R. Liu, "Modality-invariant representation for infrared and visible image registration," *arXiv preprint arXiv:2304.05646*, 2023.

[51] C. P. Chen, H. Li, Y. Wei, T. Xia, and Y. Y. Tang, "A local contrast method for small infrared target detection," *IEEE transactions on geoscience and remote sensing*, vol. 52, no. 1, pp. 574–581, 2013.

[52] J. Han, Y. Ma, B. Zhou, F. Fan, K. Liang, and Y. Fang, "A robust infrared small target detection algorithm based on human visual system," *IEEE*

*Geoscience and Remote Sensing Letters*, vol. 11, no. 12, pp. 2168–2172, 2014.

[53] X. Shao, H. Fan, G. Lu, and J. Xu, "An improved infrared dim and small target detection algorithm based on the contrast mechanism of human visual system," *Infrared Physics & Technology*, vol. 55, no. 5, pp. 403–408, 2012.

[54] S. Qi, J. Ma, C. Tao, C. Yang, and J. Tian, "A robust directional saliency-based method for infrared small-target detection under various complex backgrounds," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 3, pp. 495–499, 2012.

[55] C. Gao, D. Meng, Y. Yang, Y. Wang, X. Zhou, and A. G. Hauptmann, "Infrared patch-image model for small target detection in a single image," *IEEE transactions on image processing*, vol. 22, no. 12, pp. 4996–5009, 2013.

[56] S. Moradi, P. Moallem, and M. F. Sabahi, "A false-alarm aware methodology to develop robust and efficient multi-scale infrared small target detection algorithm," *Infrared Physics & Technology*, vol. 89, pp. 387–397, 2018.

[57] Y. Qi and G. An, "Infrared moving targets detection based on optical flow estimation," in *Proceedings of 2011 International Conference on Computer Science and Network Technology*, vol. 4. IEEE, 2011, pp. 2452–2455.

[58] F. Zhao, T. Wang, S. Shao, E. Zhang, and G. Lin, "Infrared moving small-target detection via spatiotemporal consistency of trajectory points," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 1, pp. 122–126, 2019.

[59] Y. Lu, S. Huang, and W. Zhao, "Sparse representation based infrared small target detection via an online-learned double sparse background dictionary," *Infrared Physics & Technology*, vol. 99, pp. 14–27, 2019.

[60] S. Li, C. Li, X. Yang, K. Zhang, and J. Yin, "Infrared dim target detection method inspired by human vision system," *Optik*, vol. 206, p. 164167, 2020.

[61] Y. Cui, T. Lei, G. Chen, Y. Zhang, L. Peng, X. Hao, and G. Zhang, "Hollow side window filter with saliency prior for infrared small target detection," *IEEE Geoscience and Remote Sensing Letters*, 2023.

[62] D. Zhou and X. Wang, "Robust infrared small target detection using a novel four-leaf model," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.

[63] M. Zhao, W. Li, L. Li, A. Wang, J. Hu, and R. Tao, "Infrared small uav target detection via isolation forest," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[64] C. Yu, Y. Liu, S. Wu, Z. Hu, X. Xia, D. Lan, and X. Liu, "Infrared small target detection based on multiscale local contrast learning networks," *Infrared Physics & Technology*, vol. 123, p. 104107, 2022.

[65] Q. Hou, Z. Wang, F. Tan, Y. Zhao, H. Zheng, and W. Zhang, "Ristdnet: Robust infrared small target detection network," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[66] X. Xi, J. Wang, F. Li, and D. Li, "Irsdet: Infrared small-object detection network based on sparse-skip connection and guide maps," *Electronics*, vol. 11, no. 14, p. 2154, 2022.

[67] Y. Chen, L. Li, X. Liu, and X. Su, "A multi-task framework for infrared small target detection and segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–9, 2022.

[68] C. Kwan, D. Gribben, and B. Budavari, "Target detection and classification performance enhancement using superresolution infrared videos," *Signal & Image Processing: An International Journal (SIPIJ) Vol*, vol. 12, 2021.

[69] X. Ying, Y. Wang, L. Wang, W. Sheng, L. Liu, Z. Lin, and S. Zhou, "Local motion and contrast priors driven deep network for infrared small target superresolution," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 5480–5495, 2022.

[70] S. Yao, Q. Zhu, T. Zhang, W. Cui, and P. Yan, "Infrared image small-target detection based on improved fcos and spatio-temporal features," *Electronics*, vol. 11, no. 6, p. 933, 2022.

[71] S. Zhou, Z. Gao, and C. Xie, "Dim and small target detection based on their living environment," *Digital Signal Processing*, vol. 120, p. 103271, 2022.

[72] T. Ma, Z. Yang, J. Wang, S. Sun, X. Ren, and U. Ahmad, "Infrared small target detection network with generate label and feature mapping," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[73] Q. Meng, X. Gu, X. Xu, and F. Zhou, "Basket-based softmax," *arXiv preprint arXiv:2201.09308*, 2022.

[74] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[75] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

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

[77] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in neural information processing systems*, 2019, pp. 8026–8037.

[78] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8922–8931.

**Runze Hou** recived his B.E. degree from the School of Automation, Southeast University, Nanjing, China in 2021. He is currently working toward his M.S. degree in the Tsinghua-Berkerly Shenzhen Institute of Tsinghua University. His research interests include deep learning, multi-modal learning, infrared target detection.

**Puti Yan** received the B.S. degree in Electronical Information Science and Technology from Harbin Institute of Technology,Harbin,China in 2017 and the M.S. degree in the aerospace science and technology from the Harbin Institute of Technology, Harbin, China in 2019.He is currently pursuing aerospace science and technology,Harbin Institute of Technology.He has published research paper in IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

**Xuguang Duan** is a graduate student at the Department of Computer Science and Technology, Tsinghua University. He got his B.E degree at the Department of Electronic Engineering, Tsinghua University. His research interests include machine learning, neural-symbolic systems, video understanding. He has published some research papers in top conferences and journals include NeurIPS, ICML,TPAMI, ACM Multimedia, etc.

**Xin Wang** is currently an Assistant Professor at the Department of Computer Science and Technology, Tsinghua University. He got both of his Ph.D. and B.E degrees in Computer Science and Technology from Zhejiang University, China. He also holds a Ph.D. degree in Computing Science from Simon Fraser University, Canada. His research interests include multimedia intelligence and recommendation in social media. He has published over 100 high-quality research papers in top conferences and journals including ICML, NeurIPS, IEEE TPAMI, IEEE TKDE, ACM KDD, WWW, ACM SIGIR, ACM Multimedia etc. He is the recipient of 2020 ACM China Rising Star Award and 2022 IEEE TCMC Rising Star Award.