Information Fusion 91 (2023) 445–457

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

Shape-Former: Bridging CNN and Transformer via ShapeConv for multimodal image matching

Jiaxuan Chen [a], Xiaoxian Chen [a], Shuang Chen [a,c], Yuyan Liu [a], Yujing Rao [a,b], Yang Yang [a,b,*], Haifeng Wang [a], Dan Wu [a]

[a] The Laboratory of Pattern Recognition and Artificial Intelligence, Yunnan Normal University, Kunming, 650500, China
[b] The School of Information Science and Technology, Yunnan Normal University, Kunming, 650500, China
[c] The Department of Environmental Science and Engineering, Fudan University, Shanghai, 200000, China

ARTICLE INFO

Keywords:
Feature matching
Deep learning
Shape-Former
Multimodal image matching
Registration and fusion

ABSTRACT

As with any data fusion task, the front-end of the pipeline for image fusion, aiming to collect multitudinous physical properties from multimodal images taken by different types of sensors, requires registering the overlapped content of two images via image matching. In other words, the accuracy of image matching will influence directly the subsequent fusion results. In this work, we propose a hybrid correspondence learning architecture, termed as Shape-Former, which is capable of solving matching problems such as multimodal, and multiview cases. Existing attempts have trouble capturing intricate feature interactions for seeking good correspondence, if the image pairs simultaneously suffer from geometric and radiation distortion. To address this, our key is to take advantage of convolutional neural network (CNN) and Transformer for enhancing structure consensus representation ability. Specifically, we introduce a novel ShapeConv so that CNN and Transformer can be generalized to sparse matches learning. Furthermore, we provide a robust soft estimation of outliers mechanism for filtering the response of outliers before capturing shape features. Finally, we also propose coupling multiple consensus representations to further solve the context conflict problems such as local ambiguity. Experiments with variety of datasets reveal that our Shape-Former outperforms state-of-the-art on multimodal image matching, and shows promising generalization ability to different types of image deformations.

1. Introduction

Image matching refers to correctly identifying the similar content from two or more images, which is one of the hot research topics in the area of remote sensing [1–4], medical science [5,6], and computer vision [7–9]. Especially for multimodal image fusion tasks [10–12], the latent overlapping targets have significant geometric (e.g., affine, and epipolar geometry) and radiometric differences that are typically caused by different imaging conditions such as separate sensors, and viewpoint changes. In order to yield richer scene representations via image fusion, a necessary procedure is the alignment (registration) of the various multi-sensor datas, and image matching is an important tool for image registration, where the transformation function can be fitted by a set of dependable feature matches. Thus, image matching has become a relevant area of research within the field of image fusion [13]. This paper focuses on solving sparse matching problem (i.e., following the detect-and-describe paradigm) between two sets of discrete feature points extracted by a feature detector, e.g., scale invariant feature transform (SIFT) [14].

Combinatorial nature of the matching problem leads to computational complexity ($N!$ permutations) [15]. A popular practice tackling such challenge is to consider the matching task as a two-stage mode, namely, putative set generation and outlier (false correspondences) rejection [13,16,17]. However, for the images with large nonlinear radiation distortions (i.e., multimodal image), traditional feature detection and description methods using intensity or gradient information are very sensitive to radiation distortions [18]. Fortunately, radiation-variation insensitive feature algorithms have been proposed in recent years, e.g., radiation-variation insensitive feature transform (RIFT) [19], which makes it an easy mission to construct a putative set from multimodal images. This means that the aforementioned two-stage based pipeline can be simultaneously applied to address mono-modal and multi-modal matching problems. On the other hand, the putative set is frequently contaminated by numerous outliers, which is attributed to the ambiguity of local image feature (especially when the images suffer from radiation and geometric differences) [15]. Under

* Corresponding author at: The Laboratory of Pattern Recognition and Artificial Intelligence, Yunnan Normal University, Kunming, 650500, China.
E-mail address: yyang_ynu@163.com (Y. Yang).

https://doi.org/10.1016/j.inffus.2022.10.030
Received 12 November 2021; Received in revised form 25 October 2022; Accepted 28 October 2022
Available online 3 November 2022

the circumstances, the matching task boils down to recognizing the outliers in the putative set, and how to prevent outliers from affecting downstream tasks is a long-standing problem in computer vision community [20]. The most prevalent outliers rejection methods are random sample consensus (RANSAC) [21] and its variants (e.g., MAGSAC [22]). However, if there are a lot of outliers or multiple plausible geometric models in the initial correspondence set, RANSAC-family hard to identify the potential inliers via seeking the largest subset conforming to a task-specific transformation model [23,24]. To cast off the fundamental limitations of hypothesize-and-verify technique, great efforts [4,15,25] have been spent on relaxing geometric constraint to capture local motion coherence to filter outliers. This strategy achieves better performance in scenes suffering motion discontinuities. However, underlying intrinsic geometry information of the putative set is ignored [26]. Several works recently investigated using deep learning to follow this idea, but putative correspondences generated are discrete and sparse, which makes it difficult to estimate the potential neighborhood consensus in a learning fashion. One common practice is to establish local consistent assumption in advance [2,17,27,28], so they also suffer from the similar defect of failing to capture intricate structure feature.

To address the before-mentioned issues, our primary idea is to build an end-to-end hybrid neural network, termed as Shape-Former, for learning neighborhood consensus directly from putative set. In other words, this work tries to overcome the limitations requiring predefined geometric transformation and handcrafted local consistent modeling, but leverage a same principle (i.e., deducing a correspondence as inlier or outlier hinges on local structure context information) with existing local consensus methods [2,17,27,28]. Obviously, convolutional neural network (CNN) has advantages in capturing local structure cues, but cannot handle irregular data directly. Moreover, outliers are usually distributed unevenly over the putative set, which will hinder the learning of consensus cues, for example, some outliers also have high local structure compatibility (consistency). Existing approaches usually mitigate this problem by utilizing multi-neighborhood information such as [2,15,17,27,28], but the effect is very sensitive to the setting of the multi-neighbor parameters. As a consequence, our network is designed as a two-branch architecture, where one of the branches generalizes CNN to sparse matches learning, and the other achieves a gating mechanism to obtain reliable local information. In a nutshell, our main contributions are threefold:

- We design a permutation-invariant network operation (ShapeConv) to make traditional CNN suitable for capturing local shape details of sparse data, thereby allowing us to leverage spatially-local correlation to analyze the potential neighborhood consensus across two feature point sets.
- We customize a Shape-Former network, which inherits the superiorities of both CNN and Transformer in local and global cues modeling. This architecture can provide a robust soft estimation of outliers mechanism, thus filtering the response of outliers before capturing shape features.
- We propose to learn multiple consensus representations to mine global and local context conflicts, which can effectively alleviate ambiguity problems caused by single structure information.

2. Related works

Formally, brute-force matching followed by outlier filtration [13,29], which casts the feature correspondence task into a two-stage process. We will briefly review the relevant material in the following.

2.1. Correspondence generation

Scale-invariant feature transform (SIFT) [14] is one of the most prevailing and effective keypoint extraction and feature description methods, where keypoint is detected via constructing a Gaussian scale space and leveraging a gradient histogram to describe features. SURF [30], which accelerates this procedure via Haar wavelet calculations and integral image strategy. Another famous method is ORB [31], which utilizes FAST detector [32] and BRIEF descriptor [33] to achieve a much higher speed. For multimodal images, however, the targets to be matched have significant geometric deformations (e.g., affine, epipolar geometry, and non-rigid) and even radiometric differences that are typically caused by different imaging conditions (e.g., imaging sensors). Improvements based on existing strategies have been proposed to solve multimodal matching problems. SR-SIFT [34] is introduced for multispectral image registration, which includes a scale restriction strategy and gradient orientation modification description. Hasan et al. [35] proposed SAR-SIFT, a new gradient definition to ameliorate the robustness to speckle noise, according to the specific characteristics of synthetic aperture radar (SAR) images. Another SIFT descriptor improvement, i.e. histogram of oriented phase congruency (HOPC), was introduced in [36] by leveraging phase congruency (PC) as a proxy for gradient information to ensure the commonality between multisensor remote sensing images. Fan et al. [37] proposed a feature-based registration framework by combining PC structural descriptor and improved Harris [38] to register optical-LiDAR (light detection and ranging), IR-optical, and SAR-optical images. A more general radiation-variation insensitive method based on PC and maximum index map for multimodal matching, called RIFT, was proposed in [19]. RIFT is shown promising matching performance in images suffered nonlinear radiation distortions (such as map-optical, and depth-optical). Putative matches can be easily constructed using these off-the-shelf methods, but also inevitably contaminated by outliers. Thus, it is critical to utilize outlier filtering techniques for boosting the reliability of correspondence.

Additionally to the correspondence generation approaches mentioned above, some other matching techniques without following the two-step matching strategy, have been investigated in recent years. In other words, they do not require building putative set through local image descriptor such as SIFT, but directly learn the correspondences. Several representative studies include Correspondence Transformer (COTR) [39], and TransMatcher [40]. COTR can yield a multiscale pipeline able to supply high quality correspondences by recursively zooming. The matching procedure of COTR is feeding an image pair and a query point in one image, and then finding its correspondence in another image. TransMatcher proposed a new simplified decoder keeping only the query-key similarity computation, and the matching result is decoded by the global max pooling and a MLP head. This tactic achieves satisfying performance in person re-identification. However, these direct matching methods usually require an outlier removal technique to improve robustness in challenging tasks.

2.2. Outlier rejection

Resampling technique, represented by the well-known RANSAC [21], is a prevalent geometric verification paradigm. Basically, the two keypoint sets extracted from the images are assumed to be coupled by a certain parametric geometric transformation, then RANSAC estimates a given parametric model (e.g., epipolar geometry) as hypothesis by repeatedly sampling a minimal subset, and finally verifies the confidence by the number of consistent inliers. Therefore, RANSAC and its variants (e.g., LO-RANSAC [41], GroupSAC [42], and PROSAC [43]) are also known as hypothesize-and-verify approaches. More recently, Barath et al. [22] applied σ-consensus in their MAGSAC, to eliminate the need of a pre-defined threshold by marginalizing over a range of noise scales. Its improved version (MAGSAC++) [44], which can be formulated as an M-estimation solved by the iteratively re-weighted least squares, is also introduced to avoid requiring the inlier–outlier decision. Certain fundamental drawbacks are exhibited by the RANSAC-family despite being widely used as geometric estimator. The minimal subset sampling mechanism only applies to parametric transformation

constraints, which means that resampling frameworks fail to address complex matching patterns including non-rigid deformation. In addition, the performance of RANSAC-family deteriorates dramatically with the decrease of inlier rate due to the fact that the sampled subset is prone to including outliers inevitably. This has been proven in the literatures [29,45]. On the other hand, with the runaway success of deep learning, trainable fundamental matrix estimation methods [46,47] are also started to dabble in development. Although they demonstrate promising performance, but the predicted model is not as good as using naive RANSAC in specific scenarios.

Deep learning has gained great success on complex computer vision tasks in different contexts (e.g., graph matching [48,49], and stereo matching [50]) in recent years, which motivates scholars to figure out the matching problem via learning technique. Traditional CNNs, however, is infeasible for correspondence learning due to the unordered and dispersed nature of match data. Solid point learning framework is lacking until a few years ago. PointNet [51] and its improved version PointNet++ [52] have triggered the upsurge of correspondence learning. Yi et al. [53] first attempted to introduce a context normalization (CN) in PointNet for addressing the correspondence pruning, namely, training an MLP from sparse matches together with the image intrinsics under geometrical transformation constraints. Nevertheless, using a simple normalization operation overlooks the underlying complex relations, and may hinder the overall performance. To this end, Zhao et al. [54] expected to extract reliable local information based on their NM-Net, instead of spatially nearest information, where the proposed compatibility-specific neighbor mining relying on affine attributes plays a crucial role. Zhang et al. [55] designed an OANet to cluster input correspondences by DiffPool layers, and perform full size prediction using unpooling techniques. Learning directly from matches data via MLP-based framework is a powerful solution, since MLPs are good at modeling long-distance feature dependencies of unordered sparse correspondences. However, PointNet-like frameworks still suffer from the dominant outliers included in the input correspondences, and experience difficulty to perceiving local shape details compared with the CNNs. Furthermore, most of them may sacrifice considerable inliers to estimate the motion parameters, thus limiting their application scenarios [29]. Most recently, Zhong et al. [56] proposed a permutation-equivariant split attention network (PESA-Net) to improve the ability of existing PointNet-like framework in context information learning, by an attention module including split, squeeze, excitation and union operation. Specifically, PESA-Net leverages the concepts of attention mechanism in CNN [57], and further introduces multiple learning paths, thus gathering the channel-wise contextual information and bringing performance gain in two-view correspondences. However, it is still based on the MLP framework, and requires predicting the relative pose by essential matrix. In our method, by contrast, the contextual structure features are extracted by CNN, and Shape-Former also provides a Transformer branch to filter outliers for mitigating context conflicts.

The motivation of local consensus stems from a simple physics rule, namely, the local distribution of putative set must adhere to certain local smoothing constraints, thus most feature points would keep the topological structure of their neighboring point pairs after transformation. Grid-based motion statistics (GMS) [25,58] is a statistic framework incorporated the smoothness constraint for seeking inliers. Due to discarding the structural context, GMS will be severely degraded under large view changes. To address this, locality preserving matching (LPM) [15] makes more restrictive assumptions than GMS, namely, considering to preserve the local neighborhood topologies of potential inliers. Analogously, Li et al. [59] also designed a descriptor describing the structure information of feature points, called support-line, but considers both geometric constraints and photometric. The key of these representative studies can be summarized as concluding whether a correspondence is correct or not by gathering supporting evidence from the correspondence-level local information. However, existing handcrafted structure metrics are shown to be sensitive to rotation, wide baseline, etc. From a new perspective, Jiang et al. [23] casted outlier rejection as a spatial clustering problem, which leveraged classical clustering method (DBSCAN) [60] to adaptively seek several motion consistent clusters. The fluctuation of density and the variation of the distance between clusters, nonetheless, are still the tricky matters.

The research of local consensus is also experimenting to introduce deep learning, but due to the irregular and unordered matches data, designing a handcrafted structure representation and then feeding to a neural network is popular practice. Ma et al. [17] proposed learning for mismatch removal (LMR) by using handcrafted representation and neural network. The key idea is to exploit the consensus of neighborhood elements and topology, which are fed into an inlier/outlier classifier. Chen et al. [27] introduced a self-attention MLP-based framework and a redesigned rotation invariance descriptor for learning neighborhood consistency. This strategy for relaxing constraints reveals significant superiority in dealing with general matching problems, but also inherits the inherent shortcomings of purely handicrafted approaches. For weakening handcrafted manipulations, Chen et al. [28] attempted to map the local structure of feature points directly (i.e., local structure visualization), and then input it into a customized attention network (LSV-ANet) for local consensus learning. This regular representation allows CNNs to automatically extract task-relevant consensus representations, thus averting the traditional structure measure (e.g., distances and angles). LSV-ANet improves the accuracy of matching due to the advantages of CNN in local modeling, but it is still limited to non-differentiable representation function. Overall, handcrafted techniques can easily preserve the shape information of local region around a feature point and then collect structure cues for rejecting outliers by proper network backbone. Nevertheless, they also inevitably ignore useful patterns hidden in the raw data. Technically, our Shape-Former also develops the notion of structure consensus, but we go one step further, i.e., jointly optimizing structure representation learning and local consensus evaluation from a set of putative matches.

3. Shape-former architecture

In this work, we combine the global interaction of Transformer with the power of CNN at local processing by a novel point set operation (i.e. ShapeConv). Furthermore, we propose to learn multiple context consensus for searching good correspondences.

3.1. Problem formulation

Given an image pair (I, I'), putative correspondence set S can be established via any off-the-shelf detector and descriptor, either traditional handcrafted methods or learning-based ones. Formally, let us denote the putative sparse matches as $S = [(x_1, y_1, x'_1, y'_1), \dots, (x_N, y_N, x'_N, y'_N)]$, where (x_i, y_i) and (x'_i, y'_i) indicates the 2-D keypoints located in I and I', respectively. For brevity, putative match (x_i, y_i, x'_i, y'_i) is written p_i. Normally, putative set S contains a huge proportion of outliers, and our work is to build an end-to-end hybrid neural network (i.e., Shape-Former) to overcome the limitations of existing local consensus methods. The overview of our workflow is demonstrated in Fig. 1(a). Typically, consensus technique predict the probabilities of putative matches being inliers based on local information similarity, as illustrated in Fig. 1(b). For example, adopting neural network to cast the correspondence selection as an inlier/outlier classification

$$\mathcal{Z}_i = \mathrm{Softmax}(f(R(p_i; S))), \tag{1}$$

where $f(\cdot)$ is a neural network model, $R(\cdot)$ denotes a manually engineered criteria such as geometric description including distances and angles [17,27], or grid data representing the local topological structures [2,28]. Obviously, the key here is to construct a proper

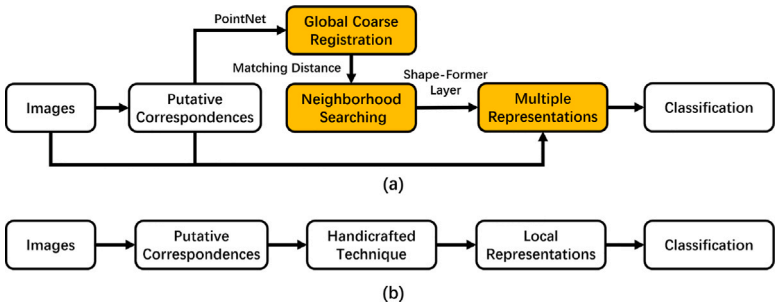

Fig. 1. (a) The flowchart of the proposed outlier rejection method. The crucial steps are marked in orange. (b) The traditional local consensus pipeline. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

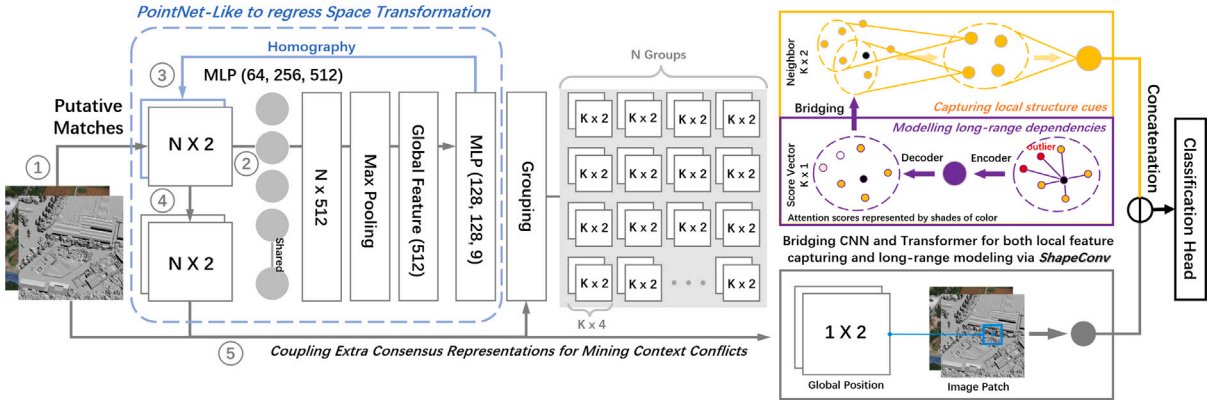

Fig. 2. The overall framework of Shape-Former for learning good correspondences. First, at the front end of the network, PointNet-like blocks are used to evaluate a projection transformation for coarse registration. Next, putative correspondences $S \in \mathbb{R}^{N\times 4}$, including two sets of feature points of size $N \times 2$, are divided into N groups of neighbors of size $K \times 4$ by a matching distance, then coupling multiple representations for each neighbor. Finally, a CNN head is used to perform inlier/outlier classification by mining consensus information. Note that the numerical labels indicate the order of computation.

structure representation, but it is difficult to ensure the validity of hand-crafted representation for different matching patterns. Therefore, our work advocates for giving network the ability to actively embedding structure context, thus avoiding manually designing complicated local measurement rules $R(\cdot)$. On the one hand, 2D convolution operations can effectively capture explicit structural information, but the typical convolution is ill-suited for handling irregular and unordered data representation, which will result in variance to ordering and deserting of shape information. On the other hand, we expect the proposed Shape-Former can effectively learn both spatially-local correlations and long-range interactions directly from sparse matches for enhancing representational ability. Thus, we introduce ShapeConv as bonder to integrate CNN with Transformer. In addition, at the front end of the network, we predict a relative transformation from S to alleviate the absolute distance changing significantly due to viewpoint changes. The Shape-Former can be expressed as:

$$\begin{aligned} &\mathcal{T} = \mathbf{F}_t(S),\ \{\mathcal{P}_i\}_{i=1}^{N} = \mathbf{F}_g(S;\mathcal{T},f_d),\\ &SC_i^{\ell} = \mathrm{ShapeConv}_{\ell}(\mathcal{P}_i;\mathrm{Fomer}_{\ell}(\mathcal{P}_i)),\ \ell = 1,\dots,L,\\ &\mathcal{Z}_i = \mathrm{CNN}(Concat(SC_i^1,\dots,SC_i^L,\mathcal{AC}_i)). \end{aligned} \tag{2}$$

$\mathbf{F}_t$ is a PointNet-based geometric estimator (with input S) to fit a homography matrix $\mathcal{T}$ for global coarse registration. $\mathbf{F}_g$ denotes a correspondence grouping rule, which allocates neighbor $\mathcal{P}_i \in \mathbb{R}^{K\times 4}$ (K is the number of neighborhood elements), based on matching distance mapping f_d, to per putative match p_i. The pivotal role of global-to-local strategy lies in considering both rigid and non-rigid deformation, thus making the proposed framework more suitable for different applications. L Shape-Former layers, stacked by ShapeConv and Transformer, explicitly model the contextual structure information SC_i^{ℓ} of putative match p_i. Note that correspondence neighbor $\mathcal{P}_i$ is the input to each Shape-Former layer. Moreover, we also leverage additional representation $\mathcal{AC}_i$ to mine for context conflicts, i.e., mitigating the harmful effect of outliers. Finally, for seeking good correspondences, a CNN is used to learn multiple representation consensus. An overview of Shape-Former is presented in Fig. 2. We will introduce this architecture minutely in the following subsections.

3.2. Global coarse registration layer

For putative correspondences S, the distance between any feature correspondence may alter significantly due to large viewpoint change, which often involves projection transformation. Neural network is still sensitive to spatial variations. Therefore, similar to spatial transformer network [61], we design a coarse registration layer, at the front-end of Shape-Former, to generate homography matrix $\mathcal{T}$ assigning correspondences between two sets of feature points (see Fig. 2 left). Though not a recent invention, a cornucopia of PointNet-based have been demonstrated promising results on tasks ranging from object detection [62], registration [63] to semantic segmentation [64]. So, we adopt PointNet as our backbone. Concretely, input S is passed to a shared-MLP [64, 256, 512], and a symmetric pooling function, which aggregates global signature from all pair-wise matches. One other MLP [128, 128, 9] including a regression layer takes the input global feature, and regresses projective parameters $\theta \in \mathbb{R}^{3\times 3}$.

3.3. ShapeConv: Bridging CNN and transformer

Convolution has the unique advantages of local representation modeling, but it is difficult to apply conventional CNNs on the unordered input correspondences. On the other hand, input correspondences contain a lot of outliers, that hampers further consensus cues evaluation.

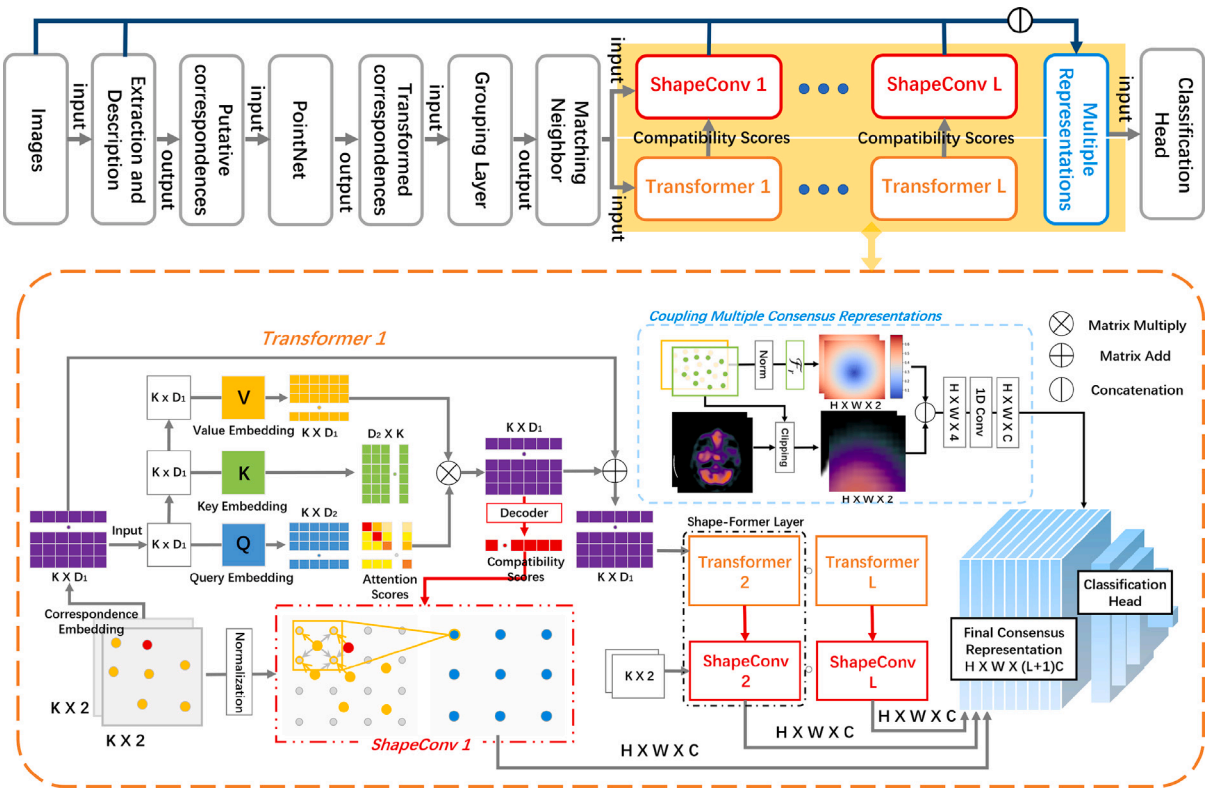

Fig. 3. Illustration about integrating CNN and Transformer via the proposed ShapeConv. Upper: the position of the proposed ShapeConv in the Shape-Former pipeline. Lower: in ShapeConv 1, • represents outlier, and the *orange arrow* represents the flow direction of the feature, which is determined by the distribution of point set. Multiple consensus representations include a global position information with image patch to address conflict of global context. We use Eq. (5) for position embedding to ensure the consistency of representation dimension, so visually it looks like a Gaussian kernel. Moreover, the match scores provided by the Transformer branch can adaptively control the feature flow direction to suppress outliers. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

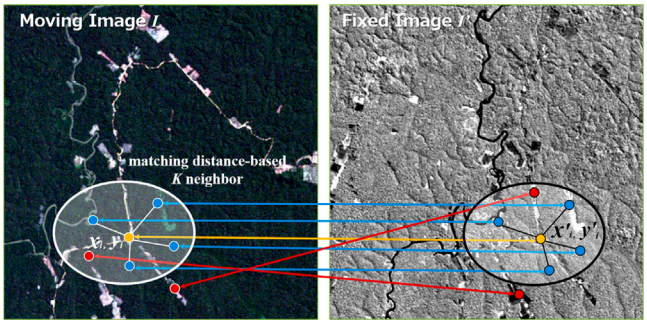

Fig. 4. Schematic illustration of the matching distance for searching K nearest neighbor of putative match p_i. *Orange arrow* represents p_i, *blue arrow* denotes the neighbor element $p_j = (x_j, y_j, x'_j, y'_j)$ that $f_d(p_j, p_i) < 5$, and *red arrows* are $f_d(p_j, p_i) \geq 5$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Therefore, the ShapeConv is designed for popularizing the typical convolution to correspondences learning, and bridging the merit of Transformer in long-range information for restraining the interference of outliers in feature extraction, as illustrated in Fig. 3.

Correspondence Grouping. It is natural to use Euclidean distance to search the neighborhood of a point. However, there is no obvious distance metric definition for sparse correspondence. Although we can use nearest neighbor strategy on one point set, and then find corresponding points in another set, but which is no guarantee that the spatially k-nearest neighbors of correspondences are coherent. To address this issue, we present a specific matching distance via a piecewise constant function in bilateral domain, to search for consistent neighbors.

Table 1
Comparison between naive nearest neighbor search and the matching distance based search. The values denote the average inlier ratio in neighbor. Dozens of image pairs are sampled from the experimental datasets (Section 4).

Neighbor	$K = 10$	$K = 15$	$K = 20$	$K = 25$	$K = 30$
NN	39.1%	38.6%	30.1%	22.1%	21.2%
Ours	67.9%	65.8%	63.2%	56.4%	38.2%

Formally, the distance between p_i and p_j can be formulized as:

$$
\begin{aligned}
&f_d(p_i, p_j) = Max\left\{ |Q_{ij}|, |Q'_{ij}| \right\}, \\
&Q_{ij} = \left\{ (x_k, y_k) \middle| d\left((x_k, y_k), (x_i, y_i)\right) < d\left((x_j, y_j), (x_i, y_i)\right) \right\}, \\
&Q'_{ij} = \left\{ (x'_k, y'_k) \middle| d\left((x'_k, y'_k), (x'_i, y'_i)\right) < d\left((x'_j, y'_j), (x'_i, y'_i)\right) \right\},
\end{aligned} \tag{3}
$$

where $|\cdot|$ indicates the cardinality for a set, and $d(\cdot)$ returns distance between two points. Note that Eq. (3) is not a mathematically strict distance metric, but we can find local matches that are within a specific radius K for a query match. For instance, the neighbors of putative match p_i can be written as:

$$
\mathcal{P}_i = \{ p_j | f_d(p_i, p_j) < K \}. \tag{4}
$$

The neighborhood elements found by matching distance can ensure the consensus of the matching vectors in the local region due to simultaneously considering the spatial relationship of $[(x_i, y_i), (x_j, y_j)]$ and $[(x'_i, y'_i), (x'_j, y'_j)]$, as shown in Fig. 4. The superiority of neighbor based on aforesaid matching distance can be further justified by the statistics, as reported in the Table 1.

Capturing Shape Representations. Feature point has a unique coordinates representing its spatial position. However, such discrete data representation does not conform to grid-based convolution kernel

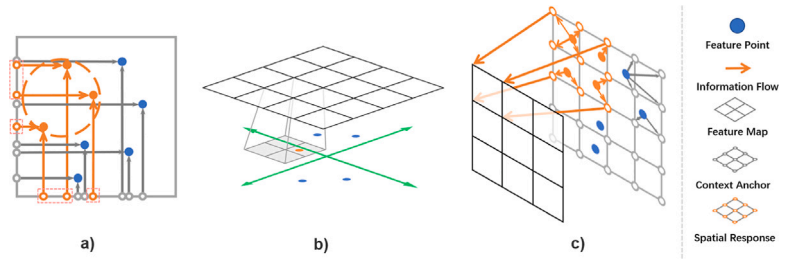

Fig. 5. (a) Point data is irregular and unordered, the coordinates of spatially adjacent points are randomly distributed on the coordinate axis. (b) Spatial distribution pattern is a latent property for point set, and the shapeConv is designed to extend the traditional convolution for leveraging spatially-local correlations of the point data. (c) ShapeConv: we leverage grid-based context anchors as sampling target, and generate spatial location response in a dense representation, where spatial response is determined by the distance between the anchor point and the nearest feature point.

for learning, as shown in Fig. 5a. On the other hand, the spatial distribution of point sets is an inherent property in Euclidean space, and it is not affected by the input order of data. This latent permutation-invariance information is easily captured via regular spatial sampling (see Fig. 5b). To solve the ill-suited problem between irregular data and grid sampling, our idea is seeking a differentiable spatial intensity response calculation mapping $\mathcal{F}_r$ for sampling the dense spatial information of a given set of points. Specifically, given a point, its spatial response at any location is defined as the distance metric of two points. Next, we leverage auxiliary grid-based anchors as sampling target, and generate spatial location response for each feature point in a dense representation. Finally, the shape representation of a point set is determined by pixel-wise minimum pooling, which can be formulated as:

$$\mathcal{F}_r(\mathcal{P}_i)_{[h,w]} = \Big(\underset{(x_j,y_j)\in\mathcal{P}_i}{\mathrm{MIN}} \Big\{ \|\mathrm{Norm}(x_j,y_j) - (h/H, w/W)\|_2 \Big\}, \underset{(x'_j,y'_j)\in\mathcal{P}_i}{\mathrm{MIN}} \Big\{ \|\mathrm{Norm}(x'_j,y'_j) - (h/H, w/W)\|_2 \Big\} \Big), \tag{5}$$

where operator $[\cdot,\cdot]$ returns the element at position (h,w), $h \in [1,H]$, $w \in [1,W]$, and $\mathrm{Norm}(\cdot)$ is normalization function. Note that H and W are the hyper-parameters determining the output size. The intuition behind spatial response function is the following: we firstly normalize the spatial coordinates of matches in $\mathcal{P}_i$ to (0, 1), so that the sampling target $(h/H, w/W)$ can overlay on each grouping. We then select Euclidean metric to calculate spatial response between all feature points and coordinates of the grid (the output size is $K\times H\times W$), and the minimum response value for each position (h,w) is reserved. This means that the final $H\times W$ representation not only guarantees permutation-invariance, but also extracts the spatial interaction features of the entire point set. Obviously, Eq. (5) is differentiable, and we can control the dimensions of the output map, thus making it easy to generalize 2D convolution operations to sparse data learning:

$$ShapeConv(\mathcal{P}_i, O)_{[h,w]} = \sum_{(n,m)\in U} \mathcal{F}_r(\mathcal{P}_i)_{[h+n,w+m]} O_{[n,m]}, \tag{6}$$

where O denotes the trainable convolution kernels within the local region U. This mechanism is analogous to describing points structure based on handcrafted descriptor in classical local consensus approaches [2,17,27,28], but allows us to learn shape features directly from raw data. Hence, we call the mechanism (Eq. (6)) point shape convolution (ShapeConv), reported in Fig. 5c. Although we directly perform structural feature extraction on the neighbor of correspondence via ShapeConv, this is not fixed local sampling. This is due to our grouping method infers geometric transformation and determines neighborhoods based on the closest matching distance, which is similar to deformable convolution [65] exploiting learnable sampling rules. However, deformable convolution leverages additional offsets, and does not work directly on irregular data.

Modeling Long-Dependencies. With the ShapeConv definition (Eq. (6)), we can follow the common practice of collecting local features in a hierarchical fashion, thus retrieving the potential inliers from putative match set S based on structural consistency cues. However, many local regions around true matches will be contaminated due to irregular distribution of outliers, which will start a chicken and egg problem: learning good structural consistency cues requires the support of potential inliers, but in the meantime, seeking good correspondences also relies on the results of reliable feature cues. A reasonable tactic is that allowing the model to jointly attend to information from different correspondences (long-distance modeling), thus estimating the degree of correlation between correspondences with the same semantics. This target can be boiled down to giving more important attention to the potential inliers, and progressively adjusting the focus with different network layers. More importantly, this soft evaluation must be used to guide local feature extraction, therefore, we design a two-branch hierarchical architecture, one of branches is Transformer, to make the ShapeConv inherit the merits of both local feature and long-distance modeling.

Like patch embedding in ViT [66], we start by a naive embedding to place putative matches having more semantically similar (i.e., correctness of correspondence) if they are spatially closer. Specifically, each putative match p_j in $\mathcal{P}_i$ is mapped to a D_1-dimensional space by a shared MLP [64, 128] with ReLU, where $D_1 = 128$ is empirically set to a relatively small value. Then, following the terminology in [67], suppose $E_{in} \in \mathbb{R}^{K\times D_1}$ denotes the input features, the query, key, and value can be defined as $\mathbf{Q} = E_{in}M_q$, $\mathbf{K} = E_{in}M_k$, and $\mathbf{V} = E_{in}M_v$, where $M_q \in \mathbb{R}^{D_1\times D_2}$, $M_k \in \mathbb{R}^{D_1\times D_2}$, and $M_v \in \mathbb{R}^{D_1\times D_1}$ denote the linear transformation matrices. In this paper, we set $D_2 = D_1/2$ for computational efficiency. Finally, we calculate the attention weighting matrix, and regression a set of correspondence compatibility scores via an MLP, since we are only interested in the points that make up the putative matches. Such score vector is used to recalibrate the spatial responses, i.e., the Eq. (5) can be rewritten as:

$$\mathcal{F}_r(\mathcal{P}_i)_{[h,w,1]} = \underset{(x_j,y_j)\in\mathcal{P}_i}{\mathrm{MIN}} \Big\{ \mathrm{MLP}(\mathrm{Softmax}(\mathbf{QK}^T/\sqrt{d})\mathbf{V} + \mathbf{V}')_{[j]} \|\mathrm{Norm}(x_j,y_j) - (h/H, w/W)\|_2 \Big\}, \tag{7}$$

in which $\sqrt{d}$ denotes an approximate normalization, and $\mathbf{V}'$ is a residual connection. For simplicity, the concept of multi-head attention here is omitted. Intuitively, here we leverage an MLP to decode match-wise dependency relationship to simplify the decoder in original Transformer. Then, treating this score vector as weights for spatial responses can increase the response value of outliers, thus filtering out the contamination of outliers by minimum pooling. This procedure shares some similarities with other CNN-based attention mechanisms such as squeeze-and-excitation network (SENet) [57] and convolutional block attention module (CBAM) [68] by capturing long-range dependencies for adaptive feature refinement. However, in our method, there is no coupling between detail capturing and long-range modeling due to two branch design. This means that the ShapeConv has a global receptive field from the get-go without deeper architectures needed. Therefore, ShapeConv are endowed with the capability of local representation and global modeling.

3.4. Coupling multiple consensus representations

Single spatial local relationship among feature points will cause ambiguities and non-deterministic conditions. Existing attempts [1,2,17,27,28] commonly use multiscale neighborhood to solve such issues. However, since it is difficult to determine an optimal combination of neighborhoods, this solution is sensitive to different matching patterns. In the image matching task, we observe that some regularities could be leveraged: (i) for a correspondence p_i, if it is an inlier, the global position of two corresponding feature points should be similar after coarse registration of the two scenes, as shown in Fig. 6 upper; (ii) even the putative match p_i has highly compatible neighborhood structure and global context, it may be wrong due to the limitation of feature description, but the image texture around it will also be different, as illustrated in Fig. 6 lower. Therefore, in this work, we propose coupling multiple consensus representations for each correspondence, thus guiding the search of good correspondences.

Clearly, the coordinates of feature points in the original image are actually the position information, but we must realize that the feature dimensionalities of ShapeConv and the coordinates of feature points are inconsistent. The ShapeConv feature maps have the dimensionality $W \times H \times C$, while the shape of the coordinates of feature point is 2×1. To address conflict of global context, we must transform global spatial information into dense representation before feeding them to a classification backbone. Fortunately, Eq. (5) can be easily used to generate the global position map of a single point. Next, we clip the local pixel blocks $\{\mathcal{IG}_i\}_{i=1}^{N}$ around each feature point to identify local conflicts, where the size of pixel block is $H \times W$. In addition, this image texture information can also help to identify global conflicts. The final classification head is implemented by a CNN, and multiple representation $\mathcal{MP}_i$ for p_i is

$$\mathcal{MP}_i = \left[g(\mathcal{F}_r(\hat{p}_i) \| \mathcal{IG}_i) \| \mathcal{SC}_i^1 \| \cdots \| \mathcal{SC}_i^L \right] \in \mathbb{R}^{H \times W \times (L+1)C}, \tag{8}$$

where $\|$ is concatenation operator, $\hat{p}_i$ denote the normalized spatial coordinates according to the entire feature set, $g(\cdot)$ is a convolution layer with C 1-D convolution kernels, and $\mathcal{SC}_i^L \in \mathbb{R}^{H \times W \times C}$ denotes the output of L Shape-Former layers.

3.5. Learning for rejecting outliers

Learning-based outlier rejection methods generally combine a binary classification loss, and have been proven effective. Therefore, our Shape-Former also uses cross entropy as classification loss, but contains two consensus constraints:

$$\mathcal{L}(S|\gamma) = \mathcal{L}_{cls}(S|\gamma) + \alpha \mathcal{L}_{sc}(S|\gamma) + \beta \mathcal{L}_{reg}(S|\gamma) \tag{9}$$

where γ denotes the parameters of Shape-Former, S is the putative set, α and β are used as trade-off parameters.

$\mathcal{L}_{cls}$ denotes native cross entropy aggregated over each individual correspondence, which can be written as:

$$\mathcal{L}_{cls} = -\frac{1}{N} \sum_{i=1}^{N} \Big[t_i \cdot log(\mathcal{Z}_i) + (1 - t_i) \cdot log(1 - \mathcal{Z}_i) \Big], \tag{10}$$

where $t_i \in \{0, 1\}$ is the groundtruth label of correspondence p_i. Specifically, 1 = inlier, 0 = outlier, and $\mathcal{Z}_i$ denotes the output of the last layer (softmax) of network for the ith putative correspondence. $\mathcal{L}_{sc}$ is a structure contrast loss, which can be viewed as a constraint term that guarantees the structure consensus of inliers and penalizes the structure correlation of negative samples during the training process. $\mathcal{L}_{sc}$ can be defined as:

$$\begin{aligned} \mathcal{L}_{sc} = \frac{1}{N} \sum_{i=1}^{N} \Big[& t_i \| \mathcal{F}_r(\mathcal{P}_i)_{[:,:,1]} - \mathcal{F}_r(\mathcal{P}_i)_{[:,:,2]} \|_F \\ & + \mathrm{MAX}\{0, (1 - t_i)(J - \| \mathcal{F}_r(\mathcal{P}_i)_{[:,:,1]} - \mathcal{F}_r(\mathcal{P}_i)_{[:,:,2]} \|_F)\} \Big] . \end{aligned} \tag{11}$$

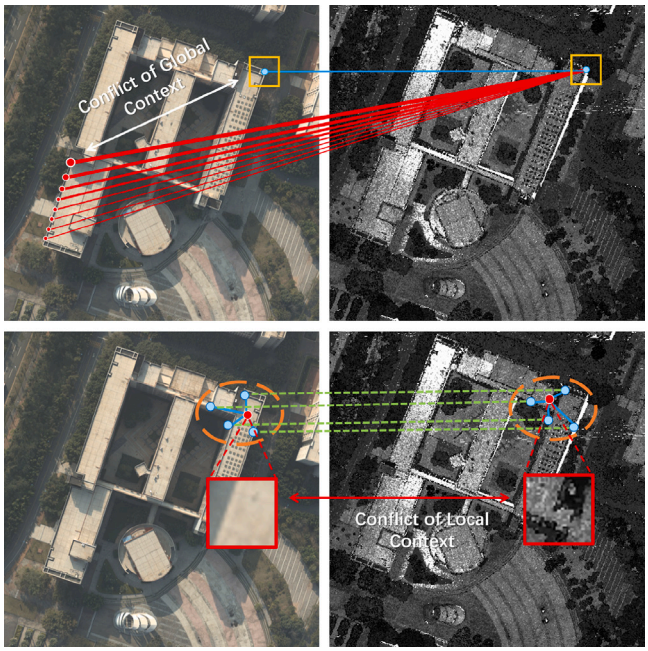

Fig. 6. Upper (conflict of global context): A feature point can have at most one correspondence in the another image. However, due to the limitation of feature description, it is easy to cause one-to-many problem. It is difficult to eliminate by structure consensus, but easy to solve with mining the conflict of global context. Lower (conflict of local context): The putative match p_i has similar neighborhood structures, but it is actually an outlier. On the other hand, similar neighborhood structures mean that these two local regions are most likely identical scenes of the objective world. Therefore, the spatial dislocation between (x_i, y_i) and (x_i', y_i') will cause significant changes in texture information around them.

In the naive contrastive loss, J is a fixed hyperparameter, but here we set $J = \mathrm{MAX}\{t_i \| \mathcal{F}_r(\mathcal{P}_i)_{[:,:,1]} - \mathcal{F}_r(\mathcal{P}_i)_{[:,:,2]} \|_F\}_{i=1}^{N}$, namely, the penalty margin is the minimum spatial response distance of inliers. For the $\mathcal{L}_{reg}$, we simply take inliers' Euclidean distance as registration loss defined on two point sets:

$$\mathcal{L}_{reg} = \frac{1}{N} \sum_{i=1}^{N} \Big[\sum_{p_i \in S} t_i \| \mathcal{T}(x_i, y_i) - (x_i', y_i') \|_2^2 \Big]. \tag{12}$$

Note that this work is a supervised method, although the loss function $\mathcal{L}$ includes the unsupervised items.

4. Experimental results

Here, we first benchmark the proposed Shape-Former against state-of-the-art (SOTA) techniques on image matching, and further summarize some insights into the modules of our method. Finally, we also report the performance of Shape-Former in some visual tasks.

Details of Implementation. Many classical image descriptors break down under strong radiometric differences or large baselines such as SIFT [14]. Therefore, we exploit RIFT [19] to build the putative correspondences for multimodal images, and SIFT for single-sensor images. The Shape-Former is trained using 20 multimodal image pairs and 10 pairs of single-modality images. Note that these 30 image pairs contain enough training samples (close to thirty thousand putative correspondences). In addition, Adam [69] algorithm is used to minimize the loss $\mathcal{L}$, where $\alpha = 0.2$, $\beta = 0.2$, and the learning rate is 1e-3. Finally, in this work, the configuration of Shape-Former is {[Former(128,4)-MLP(256,128,20)-SC(20,8×8, 3×3 × 16)]$_{\times 2}$, [Conv(3 × 3 × 128)]$_{\times 3}$, MLP(256,256,2)}, where []$_{\times 2}$ indicates Shape-Former layer is repeated for 2 times; Former(128,4) denotes linear layer parameter $D_1 = 128$, and utilizing 4 parallel attention heads in the Transformer; SC(20,8×8, 3×3 × 16) represents using 8 × 8 anchors,

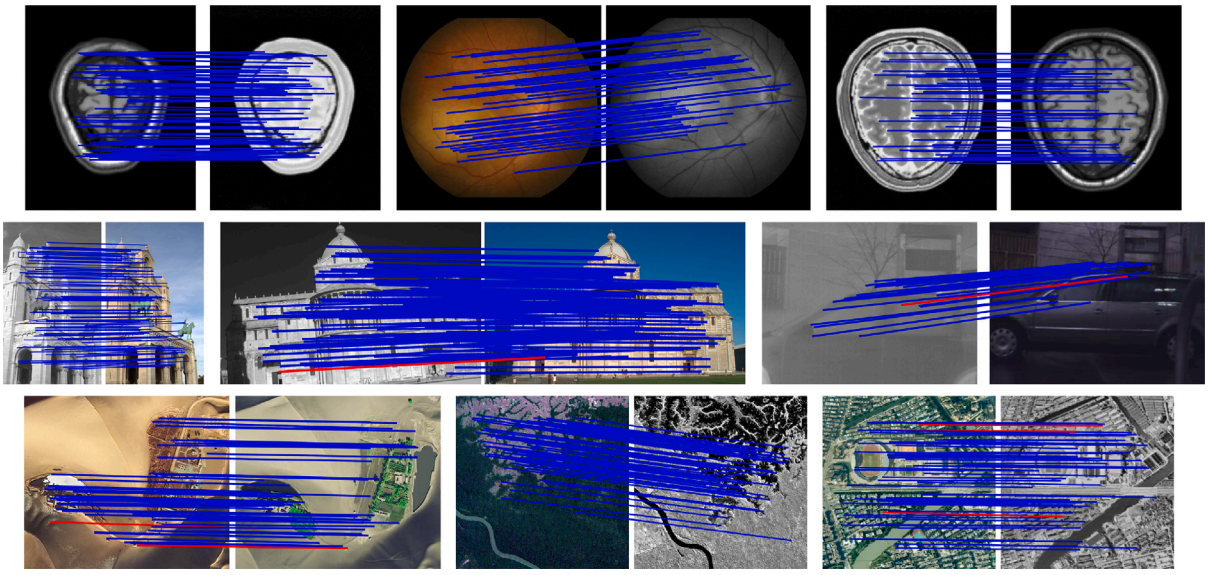

Fig. 7. Visualization results of our proposed Shape-Former on 9 typical image pairs. In each group of results, the TNs and FNs are omitted for clarity, and we show at most 200 feature matches [blue = TP, and red = FP]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3×3 convolution kernel with 16 channels, and matching distance $K = 20$ in ShapeConv. Overall, our model contains 2.40 million learnable parameters, where coarse registration network, shape-former backbone, and classification network have 0.23, 1.72, and 0.45 (millions) parameters, respectively. Several SOTA methods, including traditional handcrafted methods: RANSAC [21], MAGSAC++ [44], LPM [15], and ICF [70]; unsupervised learning method: RFM-SCAN [23]; supervised learning methods: LMR [17], OANet [55], and LSV-ANet [28], are used for comparison, and all the matching methods are implemented based on their own optimal parameter settings, and publicly available codes. For RANSAC, and MAGSAC++, we used the built-in implementation of OpenCV (4.5.3), where the transformation is homography, the upper limit of iteration is 50,000, and the pixel threshold is 5 (consistent with the groundtruth). The whole experimentation is conducted on a laptop with GeForce RTX 3060 via MATLAB and Python.

Evaluation Criterion. Following [15,28], we use precision, recall, and F-score as evaluation metrics throughout the experiments, which can be defined as:

$$\begin{aligned} \text{Recall} &= \frac{\text{TP}}{\text{TP}+\text{FN}}, \\ \text{Precision} &= \frac{\text{TP}}{\text{TP}+\text{FP}}, \\ \text{F}-\text{score} &= \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}, \end{aligned} \tag{13}$$

where TP, TN, FP, and FN mean the number of true positives, true negatives, false positives and false negatives, respectively.

Datasets. The qualitative and quantitative matching results are experimented in the following datasets.

- *Multimodal Image Matching Datasets (MIM)* [13]. *MIM* is a multimodal matching datasets, including medical, remote sensing and other computer vision (CV) multimodal images. (i) The remote sensing data contains LiDAR depth-optical, IR-optical, day-night, map-optical, SAR-optical, etc.; (ii) The medical data includes MRI-PET image pairs, SPECT-CT image pairs, T1-T2 image pairs, etc. ; (iii) The other data consists of visible-IR, visible cross-season, visible-NIR, day-night, and image-paint image pairs. Image sizes of the datasets are mostly concentrated between 256×256 and 1024×720.
- *Remote Sensing Dataset (RS)* [71]. This dataset provides color-IR, SAR, panchromatic photographs, and low-altitude SUAVs images. The images are of sizes from 256×256 to 800×600.
- *Oxford Buildings Dataset (OxBs)* [72]. *OxBs* includes vast Oxford building images, which are collected by searching for specific landmarks on Flickr. The test image pairs are resized to 800×600.

The second and third available datasets provide the putative feature correspondences and their groundtruth for each image pair. Concretely, the refined groundtruth information generated by image transformation and then manually examining the labeled correspondences. The groundtruth for the remaining datasets are established considering a benchmark as previously mentioned.

4.1. Qualitative results

First of all, we present some visualization results of our Shape-Former on several representative multi-modal image pairs (See Fig. 7). These test image pairs are collected by different imaging sensors, or suffered from different imaging conditions, including multimodal medical images (the first row of Fig. 7), multimodal outdoor images (the second row of Fig. 7), and multimodal remote sensing images (the third row of Fig. 7). From the intuitive results, we see that our Shape-Former is able to achieve promising matching performances on all the nine typical pairs.

4.2. Quantitative results

Multimodal Images. Matching for multimodal image is an important preorder procedure of multimodal fusion, and it is also quite challenging. Normally, most image pairs have low inlier percentages due to the nonlinear radiation distortions. Hence, this experiment is first provided a comprehensive quantitative comparisons with SOTA competitors on the multimodal remote sensing images and multimodal medical images. All the quantitative results for recall, precision, F-score and running time are reported in Tables 2 and 3. From the results, we can clearly see that Shape-Former is clearly superior to the other eight in F-score indicator. Generally, using resampling or parametric model (such as RANSAC and MAGSAC++) can achieve satisfying precision since most feature points are linear transformations, which are consistent with their global geometrical constraints. On the other hand, non-parametric methods such as RFM-SCAN and LPM that do not rely on global geometric models are easy get a higher recall,

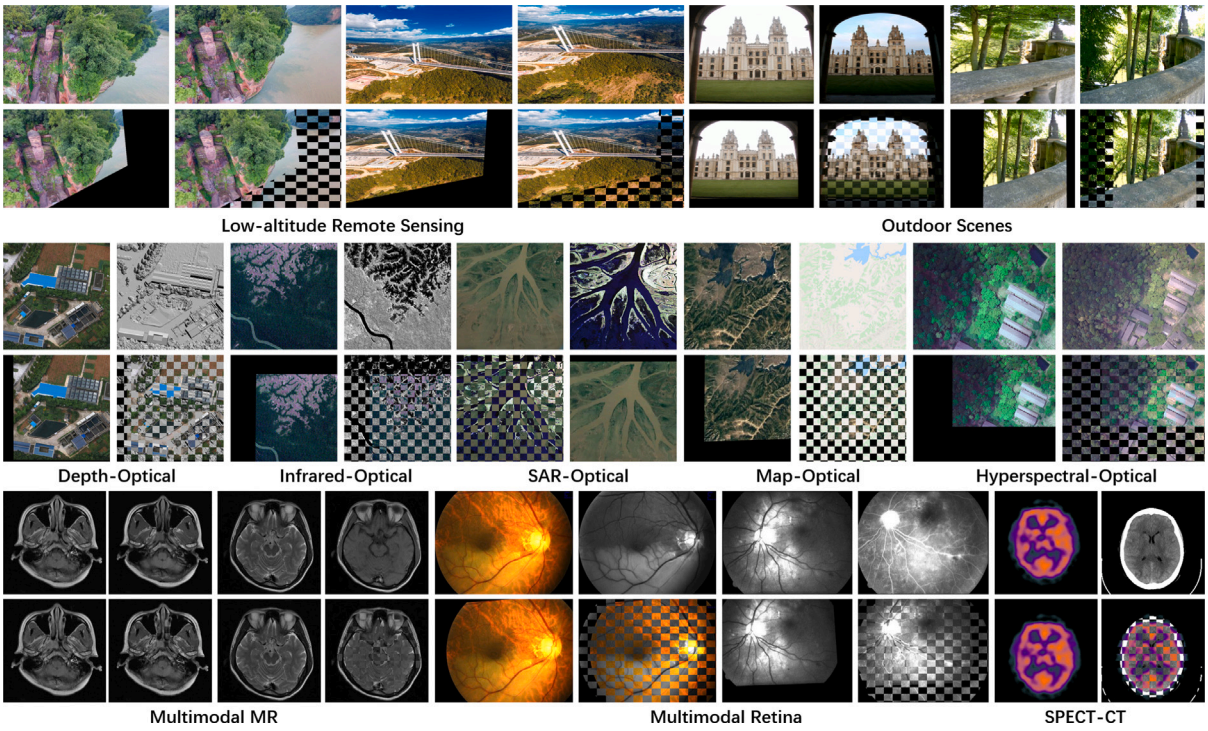

Fig. 8. Intuitive results of image registration. In each group, the first row presents moving image and fixed image, and the second row presents transformed image and checkerboard image. The warped image is obtained by sampling pixels based on bilinear interpolation.

Table 2
Quantitative comparison on MIM-RS dataset. AIR, AR, AP, AF, and AT represent average inlier ratio, average recall, average precision, average F-score, and average runtime, respectively. For clarity, **bold** indicates the best, and red font ranks the second-best.

Method/Type	Day-Night (AIR: 21.06%)				Depth-Opti. (AIR: 27.59%)				Infrared-Opti. (AIR: 46.12%)			
Metrics	AR (%)	AP (%)	AF (%)	AT (ms)	AR (%)	AP (%)	AF (%)	AT (ms)	AR (%)	AP (%)	AF (%)	AT (ms)
RANSAC	59.46	60.23	59.82	91.0	74.97	82.57	78.23	70.4	77.27	86.10	81.30	117.5
MAGSAC++	90.21	85.35	87.68	**3.8**	92.71	86.41	89.10	**2.8**	94.60	91.06	92.79	**2.3**
LPM	94.46	31.78	46.75	11.6	92.88	33.04	47.98	12.7	96.69	53.72	66.76	11.7
ICF	91.88	52.20	66.12	796.3	95.36	60.97	73.51	912.9	84.83	83.32	83.64	836.7
RFM-SCAN	**99.33**	37.83	54.36	101.7	**100.00**	38.63	54.89	119.7	**100.00**	58.22	71.36	100.6
OANet	61.53	70.03	64.81	14.2	63.35	77.92	68.49	14.3	55.12	90.80	63.48	15.0
LMR	72.40	77.75	76.81	49.2	89.42	70.47	81.01	52.4	87.99	87.61	87.11	49.4
LSV-ANet	72.58	76.57	74.64	746.1	89.50	85.87	86.15	706.5	93.29	88.24	92.69	692.7
Ours	90.67	**91.61**	**91.13**	102.8	91.35	**91.60**	**91.46**	101.1	96.29	**93.50**	**94.82**	100.7
Method/Type	Map-Opti. (AIR: 26.19%)				Opti.-Opti. (AIR: 42.57%)				SAR-Opti. (AIR: 35.16%)			
Metrics	AR (%)	AP (%)	AF (%)	AT (ms)	AR (%)	AP (%)	AF (%)	AT (ms)	AR (%)	AP (%)	AF (%)	AT (ms)
RANSAC	72.05	79.40	75.52	87.6	74.01	78.59	76.03	9.6	74.39	81.45	77.71	45.7
MAGSAC++	93.19	84.35	88.74	**1.9**	89.92	81.29	85.19	**3.0**	90.20	85.25	87.61	**2.3**
LPM	94.96	34.95	50.22	11.0	97.18	46.87	62.17	12.0	94.46	39.91	54.96	11.6
ICF	99.67	63.77	76.56	708.7	94.43	71.54	78.61	918.9	99.05	59.68	73.31	875.2
RFM-SCAN	**99.77**	46.06	62.12	91.2	**100.00**	48.64	64.36	113.2	**100.00**	44.19	60.34	99.1
OANet	61.22	80.23	66.93	14.3	43.24	72.66	50.68	13.8	59.13	72.27	64.14	13.8
LMR	89.86	84.64	85.49	46.8	84.78	80.41	81.51	50.6	82.60	85.82	80.98	47.7
LSV-ANet	91.98	79.52	86.57	689.8	92.26	85.05	87.14	610.74	91.47	89.71	88.48	644.0
Ours	93.68	**92.90**	**93.28**	100.0	96.09	**95.73**	**95.90**	99.3	94.88	**93.52**	**94.19**	99.5

but not simultaneously. For local consensus methods, LMR and LSV-ANet leverage pre-define structure representations to identify outliers, so their performance is sensitive to outliers. While OANet is a pure learning framework directly using sparse matching as input, and applies the weighted eight-point algorithm to regress essential matrix as geometric constraint. Such geometric constraint makes OANet not good at solving local deformations. On the contrary, our approach neither relies on handcrafted descriptors nor utilizes global geometric constraint.

Unisource Images. In addition, to further reveal the robustness and generalization of Shape-Former, we also consider three unisource matching scenarios: (a) satellite RS, involving only linear (e.g., affine) transformation, which commonly arises in panoramic image stitching; (b) low-altitude UAV RS for environmental monitoring, which is often used to capture the difficult terrain with the dynamic changing; (c) outdoor scenes, suffering the geometric and lighting complexity, which frequently happens in visual SLAM. The recall, precision, F-score and running time are reported in Table 4. The results show that LSV-ANet and LMR have satisfactory matching performances in all three unisource scenarios, but our Shape-Former can always produce the best precision–recall tradeoff.

Table 3
Quantitative comparison on MIM-medical dataset. AIR, AR, AP, AF, and AT represent average inlier ratio, average recall, average precision, average F-score, and average runtime, respectively. For clarity, **bold** indicates the best, and red font ranks the second-best.

Method/Type	PD-T1 (AIR: 30.02%)				PD-T2 (AIR: 27.04%)				Retina (AIR: 31.94%)			
Metrics	AR (%)	AP (%)	AF (%)	AT (ms)	AR (%)	AP (%)	AF (%)	AT (ms)	AR (%)	AP (%)	AF (%)	AT (ms)
RANSAC	19.24	19.48	19.19	4.7	11.21	10.73	10.53	4.3	73.30	75.33	74.03	40.9
MAGSAC++	78.82	74.45	76.51	**1.2**	59.70	57.98	58.67	**1.5**	89.67	81.21	85.01	**3.7**
LPM	97.34	56.61	70.94	8.9	96.62	53.00	67.57	8.9	97.64	42.22	57.99	15.9
ICF	99.84	32.43	46.76	196.6	**100.00**	30.28	43.66	232.8	98.49	63.33	75.58	2628.7
RFM-SCAN	**100.00**	54.65	69.28	33.2	**100.00**	47.11	61.36	35.8	**100.00**	39.92	56.19	316.8
OANet	59.74	**92.81**	70.98	14.1	50.76	**83.10**	60.97	14.2	48.79	66.37	54.35	14.5
LMR	53.43	76.84	65.08	35.6	71.18	76.82	71.02	35.9	88.23	85.09	85.56	71.6
LSV-ANet	70.12	65.55	66.43	700.2	73.63	74.23	73.10	638.1	89.30	83.78	85.07	975.5
Ours	88.72	88.89	**85.03**	106.5	87.66	81.53	**80.25**	103.5	96.28	**91.81**	**93.14**	101.2
Method/Type	T1-T2 (AIR: 31.23%)				RGB-NIR (AIR: 35.46%)				VIS-IR (AIR: 15.36%)			
Metrics	AR (%)	AP (%)	AF (%)	AT (ms)	AR (%)	AP (%)	AF (%)	AT (ms)	AR (%)	AP (%)	AF (%)	AT (ms)
RANSAC	22.13	20.92	21.32	5.2	68.52	71.63	70.00	29.3	6.37	6.44	6.23	37.7
MAGSAC++	69.70	68.79	69.15	**1.4**	83.86	82.45	83.14	**3.5**	36.56	32.20	34.05	**2.3**
LPM	98.16	64.71	77.69	9.5	97.33	49.92	63.86	15.0	94.01	26.75	40.70	12.5
ICF	99.95	34.59	48.94	232.6	88.92	78.82	80.14	2651.7	81.39	29.20	40.83	1583.9
RFM-SCAN	**100.00**	79.49	87.73	33.8	**99.93**	44.50	59.54	311.6	**98.65**	25.94	40.09	193.4
OANet	52.18	**97.36**	66.83	13.6	74.90	78.89	73.04	16.6	75.27	60.49	63.62	13.6
LMR	75.47	89.11	82.25	35.6	86.79	85.98	84.09	72.4	69.78	63.98	62.66	56.4
LSV-ANet	87.25	80.19	81.45	671.3	89.66	82.26	84.71	954.2	71.40	55.14	60.07	942.9
Ours	96.88	89.63	**91.41**	103.5	92.71	**91.37**	**90.74**	103.6	73.68	**75.05**	**67.12**	102.9

Table 4
Quantitative comparison on familiar single-sensor matching scenarios including high-altitude satellites images, low-altitude UAVs images and outdoor scenes. AIR, AR, AP, AF, and AT represent average inlier ratio, average recall, average precision, average F-score, and average runtime, respectively. For clarity, **bold** indicates the best, and red font ranks the second-best.

Dataset	Satellites (*RS*, AIR: 32.23%)				UAVs (*RS*, AIR: 29.62%)				Outdoor Scenes (*OxBs*, AIR: 41.23%)			
Metrics	AR (%)	AP (%)	AF (%)	AT (ms)	AR (%)	AP (%)	AF (%)	AT (ms)	AR (%)	AP (%)	AF (%)	AT (ms)
RANSAC	90.90	98.25	94.12	184.2	82.65	96.34	88.35	105.8	74.55	**98.53**	83.92	260.5
MAGSAC++	94.16	98.27	96.00	**2.6**	86.79	97.04	91.12	**1.8**	77.11	97.58	84.97	**3.6**
LPM	99.87	73.05	82.76	11.9	99.54	69.75	81.30	9.2	96.96	60.65	72.56	10.5
ICF	91.22	68.62	68.43	1550.1	76.80	74.78	66.18	444.9	73.19	75.22	65.07	206.1
RFM-SCAN	**100.0**	91.30	95.34	121.2	97.23	78.93	86.92	46.0	97.36	71.84	81.26	29.4
OANet	85.92	98.02	91.13	14.1	86.72	96.79	91.20	14.0	82.47	97.00	88.60	9.6
LMR	98.30	93.01	95.05	63.0	98.18	89.19	93.33	39.5	86.71	83.31	84.71	50.8
LSV-ANet	99.27	96.26	97.69	562.2	**99.78**	93.80	96.65	532.2	94.99	88.49	91.34	402.6
Ours	99.37	**98.68**	**99.02**	99.6	98.91	**98.42**	**98.66**	98.5	**98.48**	98.05	**98.25**	56.9

4.3. Ablation studies

To evaluate our design decisions, we perform ablation studies on multimodal image matching experiments, focusing on different Shape-Former architectures for revealing the performance gains that Shape-Former layer can deliver. The ablation studies, summarized in Table 5, shows that all modules are useful and bring substantial performance gains. Although adding more Shape-Former layers can improve matching performance, we find that this benefit will gradually weaken with the depth of Shape-Former. If more training samples are available, we expect that the matching performance can be further boosted by deeper Shape-Former architecture. Moreover, the matching distance K is a core parameter in Shape-Former, which determines the receptive field of each correspondence. Therefore, we also show how the different values of matching distance K affect the performance of our Shape-Former, as shown in Table 6.

Table 5
Ablation studies of our Shape-Former on multimodal images. MCR denotes multiple representations, Cas. stands for the number of Shape-Former layers, and Clas. represents the classification head. Note that when training more layers, we do not tune the ratios of losses and the training schedule.

MCR	Cas.	Consensus backbone	Clas.	AF (%)	AT (ms)
×	1	Shape-Former	CNN	88.12	65.6
✓	1	Shape-Former	CNN	89.26	70.5
×	2	Shape-Former	CNN	92.06	97.3
✓	2	Shape-Former	CNN	93.47	103.1
×	3	Shape-Former	CNN	93.78	125.9
✓	3	Shape-Former	CNN	94.55	126.2
×	4	Shape-Former	CNN	94.65	153.0
✓	4	Shape-Former	CNN	95.29	158.3
×	1	ShapeConv	CNN	86.68	60.1
✓	1	ShapeConv	CNN	87.47	64.3
×	2	ShapeConv	CNN	90.88	86.8
✓	2	ShapeConv	CNN	91.76	90.2
×	3	ShapeConv	CNN	92.49	112.3
✓	3	ShapeConv	CNN	93.52	118.4
×	4	ShapeConv	CNN	93.47	134.4
✓	4	ShapeConv	CNN	94.21	139.8

Table 6
The results of f-score caused by different matching distances.

Distance	$K = 10$	$K = 15$	$K = 20$	$K = 25$	$K = 30$
F-score	87.16%	92.00%	93.47%	93.53%	93.48%

4.4. Applications

To further explain the reliability of the proposed Shape-Former in practical use, we evaluate the performance of Shape-Former in image registration (fusion) and loop closure detection. In addition to the above three datasets, here, we add two datasets (SUIR [29], and KITTI [73]) for testing.

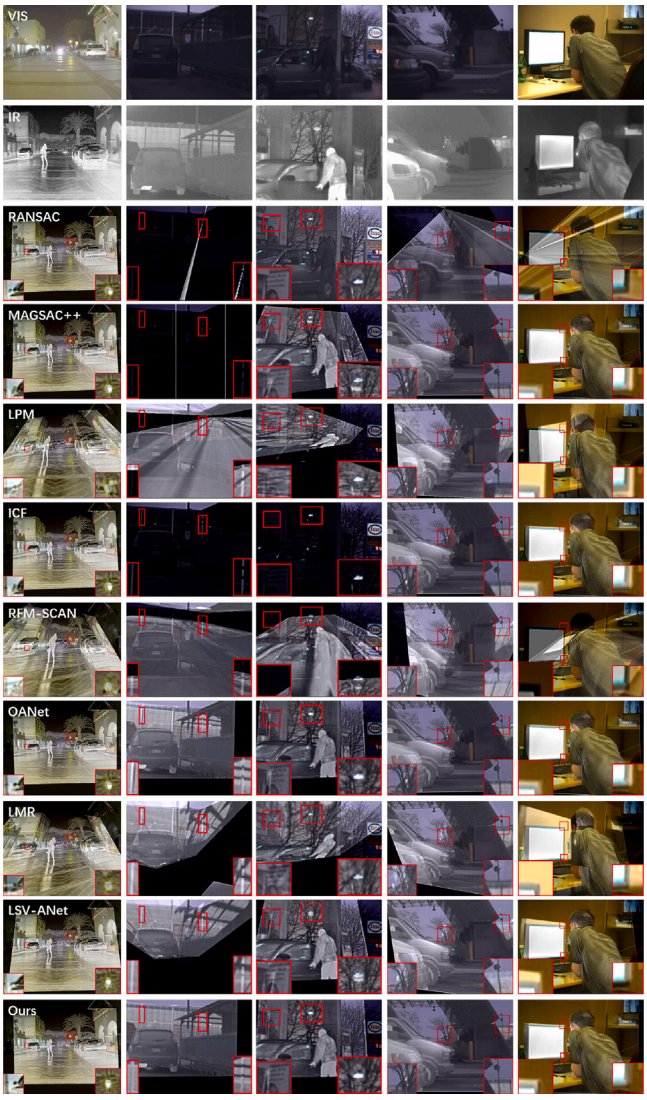

Fig. 9. The results of visible and infrared fusion. The first row presents the original visible images, the second is the infrared images, the last row is the fusion results after registration via our Shape-Former, and the rest rows correspond to eight comparison methods.

4.4.1. Image registration and fusion

Image registration and fusion is one of the important applications of feature matching. To demonstrate the effectiveness of our Shape-Former for registration and fusion, firstly extensive image registration results including multimodal RS images (e.g., SAR-optical, day-night, infrared-optical, and map-optical) and multimodal medical images (e.g., multimodal retina, multimodal MR, MR-PET, and SPECT-CT) are shown in Fig. 8. From the checkerboard diagrams, we see that Shape-Former obtained satisfactory registration accuracy. Secondly, to demonstrate the importance of our approach in subsequent image fusion, we leverage U2Fusion network [74] to fuse the infrared and visible images after registration. The qualitative results are shown in Fig. 9. We can see that the fusion results of our method have more high quality, more rich texture details, and more appropriate for human visual perception. Obviously, strict image registration helps for better fusion, and no artifacts are introduced under the visual inspection.

4.4.2. Loop closure detection

Loop closure detection, devoting the correct identification of previously observed areas, is a key component in VSLAM systems for reducing cumulative error, because this module could help amend the orientation and the estimated position. Therefore, we also apply Shape-Former to detect the loop closing pairs. For quantitative analysis, we use frame images from the KITTI (sequence 2) dataset [73] for testing. This further application can be seen as testing the robustness of the algorithm in unknown scenes. Following [29], maximum recall rate at 100% accuracy (MR) is used an essential metric, due to that more correct loop closure examples can adjust the final pose of robot and exclude accumulated error better. Quantitative results are reported in Table 7, compared with RANSAC, MAGSAC++, LPM, OANet, COTR, LMR, and LSV-ANet.

Table 7
Comparative results of MR on the KITTI dataset. **Bold** indicates the best, red ranks the second.

RANSAC	MAGSAC++	LPM	LMR
77.34	78.25	76.32	78.01
COTR	OANet	LSV-ANet	Ours
78.63	78.68	77.21	**79.16**

5. Conclusion

In this paper, we proposed a novel end-to-end outlier rejection pipeline for robust feature matching of multimodal images, which works based on a classical principle that the local structures of a true match should be similar. Without relying on global geometric constraints and structural descriptors, the proposed ShapeConv explicitly allows long-range dependencies modeling and local feature extraction of point data within a unified network, thus fully mining structure consensus for each correspondence. Experiments demonstrate that Shape-Former behaves favorably to the state-of-the-art including handcrafted and learning-based approaches, which also confirms the advantages of end-to-end consensus learning. The current shortcoming of Shape-Former is the demand for defining fixed neighborhood size. We expect to design a more advanced hierarchical structure without such limitations in the future works.

CRediT authorship contribution statement

Jiaxuan Chen: Conceptualization, Methodology, Software, Writing – original draft. **Xiaoxian Chen:** Software, Writing – review & editing. **Shuang Chen:** Visualization, Writing – review & editing. **Yuyan Liu:** Software, Validation. **Yujing Rao:** Software, Validation. **Yang Yang:** Supervision. **Haifeng Wang:** Investigation. **Dan Wu:** Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was supported by the National Natural Science Foundation of China [41971392], and the Yunnan Ten-thousand Talents Program, China.

References

[1] J. Chen, S. Chen, Y. Liu, X. Chen, Y. Yang, Y. Zhang, Robust local structure visualization for remote sensing image registration, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14 (2021) 1895–1908, http://dx.doi.org/10.1109/JSTARS.2021.3050459.

[2] J. Chen, X. Fan, S. Chen, Y. Yang, H. Bai, Robust feature matching via hierarchical local structure visualization, IEEE Geosci. Remote Sens. Lett. (2021) 1–5, http://dx.doi.org/10.1109/LGRS.2021.3099307.

[3] J. Chen, S. Chen, X. Chen, Y. Yang, Y. Rao, StateNet: Deep state learning for robust feature matching of remote sensing images, IEEE Trans. Neural Netw. Learn. Syst. (2021) 1–15, http://dx.doi.org/10.1109/TNNLS.2021.3120768.

[4] J. Ma, J. Jiang, H. Zhou, J. Zhao, X. Guo, Guided locality preserving feature matching for remote sensing image registration, IEEE Trans. Geosci. Remote Sens. 56 (8) (2018) 4435–4447, http://dx.doi.org/10.1109/TGRS.2018.2820040.

[5] G. Balakrishnan, A. Zhao, M.R. Sabuncu, J. Guttag, A.V. Dalca, An unsupervised learning model for deformable medical image registration, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2018, pp. 9252–9260.

[6] S. Zhao, Y. Dong, E.I. Chang, Y. Xu, et al., Recursive cascaded networks for unsupervised medical image registration, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2019, pp. 10600–10610.

[7] J.L. Schonberger, J.-M. Frahm, Structure-from-motion revisited, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2016, pp. 4104–4113.

[8] R. Mur-Artal, J.M.M. Montiel, J.D. Tardos, ORB-SLAM: a versatile and accurate monocular SLAM system, IEEE Trans. Robot. 31 (5) (2015) 1147–1163.

[9] A.J. Davison, I.D. Reid, N.D. Molton, O. Stasse, MonoSLAM: Real-time single camera SLAM, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 1052–1067.

[10] J. Ma, W. Yu, P. Liang, C. Li, J. Jiang, FusionGAN: A generative adversarial network for infrared and visible image fusion, Inf. Fusion 48 (2019) 11–26.

[11] H. Xu, J. Ma, EMFusion: An unsupervised enhanced medical image fusion network, Inf. Fusion (2021).

[12] Y. Ye, L. Bruzzone, J. Shan, F. Bovolo, Q. Zhu, Fast and robust matching for multimodal remote sensing image registration, IEEE Trans. Geosci. Remote Sens. 57 (11) (2019) 9059–9070, http://dx.doi.org/10.1109/TGRS.2019.2924684.

[13] X. Jiang, J. Ma, G. Xiao, Z. Shao, X. Guo, A review of multimodal image matching: Methods and applications, Inf. Fusion 73 (2021) 22–71, http://dx.doi.org/10.1016/j.inffus.2021.02.012.

[14] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.

[15] J. Ma, J. Zhao, J. Jiang, H. Zhou, X. Guo, Locality preserving matching, Int. J. Comput. Vis. 127 (5) (2019) 512–531.

[16] X. Jiang, J. Jiang, A. Fan, Z. Wang, J. Ma, Multiscale locality and rank preservation for robust feature matching of remote sensing images, IEEE Trans. Geosci. Remote Sens. 57 (9) (2019) 6462–6472, http://dx.doi.org/10.1109/TGRS.2019.2906183.

[17] J. Ma, X. Jiang, J. Jiang, J. Zhao, X. Guo, LMR: Learning a two-class classifier for mismatch removal, IEEE Trans. Image Process. 28 (8) (2019) 4045–4059, http://dx.doi.org/10.1109/TIP.2019.2906490.

[18] J. Chen, S. Chen, Y. Liu, X. Chen, X. Fan, Y. Rao, C. Zhou, Y. Yang, IGS-Net: Seeking good correspondences via interactive generative structure learning, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–13, http://dx.doi.org/10.1109/TGRS.2021.3135430.

[19] J. Li, Q. Hu, M. Ai, RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform, IEEE Trans. Image Process. 29 (2020) 3296–3310, http://dx.doi.org/10.1109/TIP.2019.2959244.

[20] J. Chen, S. Chen, X. Chen, Y. Dai, Y. Yang, CSR-Net: Learning adaptive context structure representation for robust feature correspondence, IEEE Trans. Image Process. 31 (2022) 3197–3210, http://dx.doi.org/10.1109/TIP.2022.3166284.

[21] M.A. Fischler, R.C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (6) (1981) 381–395.

[22] D. Barath, J. Matas, J. Noskova, Magsac: marginalizing sample consensus, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2019, pp. 10197–10205.

[23] X. Jiang, J. Ma, J. Jiang, X. Guo, Robust feature matching using spatial clustering with heavy outliers, IEEE Trans. Image Process. 29 (2020) 736–746.

[24] S. Chen, J. Chen, Y. Rao, X. Chen, X. Fan, H. Bai, L. Xing, C. Zhou, Y. Yang, A hierarchical consensus attention network for feature matching of remote sensing images, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–11, http://dx.doi.org/10.1109/TGRS.2022.3165222.

[25] J.W. Bian, W.Y. Lin, Y. Liu, L. Zhang, S.K. Yeung, M.M. Cheng, I. Reid, GMS: Grid-based motion statistics for fast, ultra-robust feature correspondence, Int. J. Comput. Vis. 128 (6) (2020) 1580–1593.

[26] J. Ma, Z. Li, K. Zhang, Z. Shao, G. Xiao, Robust feature matching via neighborhood manifold representation consensus, ISPRS J. Photogramm. Remote Sens. 183 (2022) 196–209.

[27] S. Chen, J. Chen, Z. Xiong, L. Xing, Y. Yang, F. Xiao, K. Yan, H. Li, Learning relaxed neighborhood consistency for feature matching, IEEE Trans. Geosci. Remote Sens. (2021) 1–13, http://dx.doi.org/10.1109/TGRS.2021.3080046.

[28] J. Chen, S. Chen, X. Chen, Y. Yang, L. Xing, X. Fan, Y. Rao, LSV-ANet: Deep learning on local structure visualization for feature matching, IEEE Trans. Geosci. Remote Sens. (2021) 1–18, http://dx.doi.org/10.1109/TGRS.2021.3062498.

[29] J. Ma, X. Jiang, A. Fan, J. Jiang, J. Yan, Image matching from handcrafted to deep features: A survey, Int. J. Comput. Vis. 129 (2021) 23–79.

[30] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (SURF), Comput. Vis. Image Underst. 110 (3) (2008) 346–359.

[31] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, ORB: An efficient alternative to SIFT or SURF, in: Proc. IEEE Int. Conf. Comput. Vis., IEEE, 2011, pp. 2564–2571.

[32] E. Rosten, T. Drummond, Machine learning for high-speed corner detection, in: Proc. Eur. Conf. Comput. Vis., Springer, 2006, pp. 430–443.

[33] M. Calonder, V. Lepetit, C. Strecha, P. Fua, Brief: Binary robust independent elementary features, in: Proc. Eur. Conf. Comput. Vis., Springer, 2010, pp. 778–792.

[34] Z. Yi, C. Zhiguo, X. Yang, Multi-spectral remote image registration based on SIFT, Electron. Lett. 44 (2) (2008) 107–108.

[35] F. Dellinger, J. Delon, Y. Gousseau, J. Michel, F. Tupin, SAR-SIFT: a SIFT-like algorithm for SAR images, IEEE Trans. Geosci. Remote Sens. 53 (1) (2014) 453–466.

[36] Y. Ye, J. Shan, L. Bruzzone, L. Shen, Robust registration of multimodal remote sensing images based on structural similarity, IEEE Trans. Geosci. Remote Sens. 55 (5) (2017) 2941–2958.

[37] B. Fan, C. Huo, C. Pan, Q. Kong, Registration of optical and SAR satellite images by exploring the spatial relationship of the improved SIFT, IEEE Geosci. Remote Sens. Lett. 10 (4) (2012) 657–661.

[38] C. Harris, M. Stephens, et al., A combined corner and edge detector, in: Alvey Vision Conference, Vol. 15, Citeseer, 1988, pp. 10–5244.

[39] W. Jiang, E. Trulls, J. Hosang, A. Tagliasacchi, K.M. Yi, Cotr: Correspondence transformer for matching across images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6207–6217.

[40] S. Liao, L. Shao, TransMatcher: Deep image matching through transformers for generalizable person re-identification, Adv. Neural Inf. Process. Syst. 34 (2021) 1992–2003.

[41] O. Chum, J. Matas, J. Kittler, Locally optimized RANSAC, in: In Proceedings of the Joint Pattern Recognition Symposium, Springer, 2003, pp. 236–243.

[42] K. Ni, H. Jin, F. Dellaert, Groupsac: Efficient consensus in the presence of groupings, in: Proc. IEEE Int. Conf. Comput. Vis., IEEE, 2009, pp. 2193–2200.

[43] O. Chum, J. Matas, Matching with PROSAC-progressive sample consensus, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2005, pp. 220–226.

[44] D. Barath, J. Noskova, M. Ivashechkin, J. Matas, MAGSAC++, a fast, reliable and accurate robust estimator, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2020, pp. 1304–1312.

[45] C. Zhao, Z. Cao, J. Yang, K. Xian, X. Li, Image feature correspondence selection: A comparative study and a new contribution, IEEE Trans. Image Process. 29 (2020) 3506–3519, http://dx.doi.org/10.1109/TIP.2019.2962678.

[46] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, C. Rother, Dsac-differentiable ransac for camera localization, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2017, pp. 6684–6692.

[47] E. Brachmann, C. Rother, Neural-guided RANSAC: Learning where to sample model hypotheses, in: Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 4322–4331.

[48] R. Wang, J. Yan, X. Yang, Combinatorial learning of robust deep graph matching: an embedding based approach, IEEE Trans. Pattern Anal. Mach. Intell. (2020) 1, http://dx.doi.org/10.1109/TPAMI.2020.3005590.

[49] R. Wang, J. Yan, X. Yang, Neural graph matching network: Learning Lawler's quadratic assignment problem with extension to hypergraph and multiple-graph matching, IEEE Trans. Pattern Anal. Mach. Intell. (2021) 1, http://dx.doi.org/10.1109/TPAMI.2021.3078053.

[50] F. Zhang, X. Qi, R. Yang, V. Prisacariu, B. Wah, P. Torr, Domain-invariant stereo matching networks, in: European Conference on Computer Vision, Springer, 2020, pp. 420–439.

[51] C.R. Qi, H. Su, K. Mo, L.J. Guibas, Pointnet: Deep learning on point sets for 3d classification and segmentation, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2017, pp. 652–660.

[52] C.R. Qi, L. Yi, H. Su, L.J. Guibas, Pointnet++: Deep hierarchical feature learning on point sets in a metric space, in: Advances in Neural Information Processing Systems, 2017, pp. 5099–5108.

[53] K. Moo Yi, E. Trulls, Y. Ono, V. Lepetit, M. Salzmann, P. Fua, Learning to find good correspondences, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2018, pp. 2666–2674.

[54] C. Zhao, Z. Cao, C. Li, X. Li, J. Yang, Nm-net: Mining reliable neighbors for robust feature correspondences, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2019, pp. 215–224.

[55] J. Zhang, D. Sun, Z. Luo, A. Yao, L. Zhou, T. Shen, Y. Chen, L. Quan, H. Liao, Learning two-view correspondences and geometry using order-aware network, in: Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 5845–5854.

[56] Z. Zhong, G. Xiao, S. Wang, L. Wei, X. Zhang, PESA-Net: Permutation-equivariant split attention network for correspondence learning, Inf. Fusion 77 (2022) 81–89.

[57] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2018, pp. 7132–7141.

[58] J. Bian, W.-Y. Lin, Y. Matsushita, S.-K. Yeung, T.-D. Nguyen, M.-M. Cheng, Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2017, pp. 4181–4190.

[59] J. Li, Q. Hu, M. Ai, R. Zhong, Robust feature matching via support-line voting and affine-invariant ratios, ISPRS J. Photogramm. Remote Sens. 132 (2017) 61–76.

[60] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al., A density-based algorithm for discovering clusters in large spatial databases with noise., in: KDD, Vol. 96, 1996, pp. 226–231.

[61] M. Jaderberg, K. Simonyan, A. Zisserman, et al., Spatial transformer networks, in: Advances in Neural Information Processing Systems, 2015, pp. 2017–2025.

[62] C.R. Qi, O. Litany, K. He, L. Guibas, Deep hough voting for 3D object detection in point clouds, in: Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 9276–9285, http://dx.doi.org/10.1109/ICCV.2019.00937.

[63] Y. Aoki, H. Goforth, R.A. Srivatsan, S. Lucey, PointNetLK: Robust amp; efficient point cloud registration using PointNet, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2019, pp. 7156–7165, http://dx.doi.org/10.1109/CVPR.2019.00733.

[64] L. Landrieu, M. Simonovsky, Large-scale point cloud semantic segmentation with superpoint graphs, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2018, pp. 4558–4567, http://dx.doi.org/10.1109/CVPR.2018.00479.

[65] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, Deformable convolutional networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2017, pp. 764–773.

[66] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021.

[67] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[68] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module, in: Proc. Eur. Conf. Comput. Vis., 2018, pp. 3–19.

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

[70] X. Li, Z. Hu, Rejecting mismatches by correspondence function, Int. J. Comput. Vis. 89 (1) (2010) 1–17.

[71] S. Zhang, W. Zhao, X. Hao, Y. Yang, C. Guan, A context-aware locality measure for inlier pool enrichment in stepwise image registration, IEEE Trans. Image Process. 29 (2020) 4281–4295, http://dx.doi.org/10.1109/TIP.2019.2961480.

[72] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, Object retrieval with large vocabularies and fast spatial matching, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2007, pp. 1–8.

[73] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? The KITTI vision benchmark suite, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2012.

[74] H. Xu, J. Ma, J. Jiang, X. Guo, H. Ling, U2fusion: A unified unsupervised image fusion network, IEEE Trans. Pattern Anal. Mach. Intell. (2020) 1, http://dx.doi.org/10.1109/TPAMI.2020.3012548.