Recurrent Homography Estimation Using Homography-Guided Image Warping and Focus Transformer

Si-Yuan Cao[1,2], Runmin Zhang[2], Lun Luo[2]*, Beinan Yu[2], Zehua Sheng[2], Junwei Li[2], Hui-Liang Shen[2]
[1]Ningbo Innovation Center, Zhejiang University, [2]College of Information Science and Electronic Engineering, Zhejiang University
karlcao@hotmail.com,{runmin_zhang,luolun,yubeinan,shengzehua,lijunwei7788,shenhl}@zju.edu.cn

Abstract

We propose the Recurrent homography estimation framework using Homography-guided image Warping and Focus transformer (FocusFormer), named RHWF. Both being appropriately absorbed into the recurrent framework, the homography-guided image warping progressively enhances the feature consistency and the attention-focusing mechanism in FocusFormer aggregates the intra-inter correspondence in a global→nonlocal→local manner. Thanks to the above strategies, RHWF ranks top in accuracy on a variety of datasets, including the challenging cross-resolution and cross-modal ones. Meanwhile, benefiting from the recurrent framework, RHWF achieves parameter efficiency despite the transformer architecture. Compared to previous state-of-the-art approaches LocalTrans and IHN, RHWF reduces the mean average corner error (MACE) by about 70% and 38.1% on the MSCOCO dataset, while saving the parameter costs by 86.5% and 24.6%. Similar to the previous works, RHWF can also be arranged in 1-scale for efficiency and 2-scale for accuracy, with the 1-scale RHWF already outperforming most of the previous methods. Source code is available at https://github.com/imdumpl78/RHWF.

1. Introduction

Homography is defined as a global projective mapping between two images captured from different perspectives. It has been widely applied in computer vision tasks ranging from the monocular camera system to the multi-camera system, such as image/video stitching [4, 17, 19], multi-scale gigapixel photography [3, 34], multispectral image fusion [41, 49], planar object tracking [44, 45], SLAM [14, 31], and GPS-denied UAV localization [18, 48].

Deep homography estimation was introduced in the pioneer [12] that uses a VGG-style network to predict the homography. Many following works have been presented to further improve the estimation accuracy, including cascading multiple similar networks [15, 21, 22, 34] or designing iterable architectures such as the IC-LK iterator [7, 48] and the trainable CNN iterator [6]. The cascading strategy has improved the accuracy to some extent but is limited by the fixed number of networks. Worse still, stacking more networks cannot guarantee better accuracy [22]. The IC-LK (inverse compositional Lucas-Kanade [1]) based deep methods use deep feature extractor combined with the untrainable iterator to improve the estimation performance, but is limited by the theoretical drawback of the untrainable iterator [6, 32]. IHN [6] avoids this limitation by designing an iterable and trainable network architecture, which further improves the estimation accuracy. However, the feature inconsistency caused by the homography deformation has long been neglected in most current works.

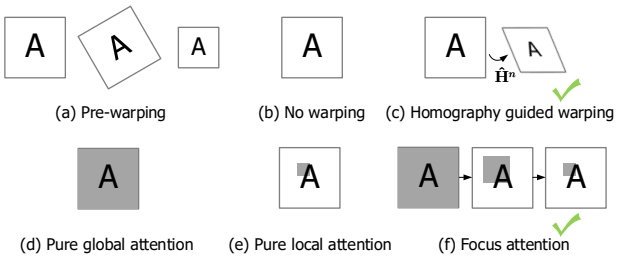


Figure 1. Illustration of the difference of warping and attention strategies in RHWF and previous approaches. Our RHWF deploys (c) and (f). Please see text for details.

It has been well investigated in [9] that standard convolution is unable to keep the equivariance under the spatial transformation except translation. However, besides translation, homography is composed of rotation, scaling, shearing, aspect ratio, and perspective transformations [37, 43], which leads to the inconsistency of the features from corresponding points [25]. The inconsistency will hinder the homography estimation performance. Many efforts have been made to acquire the transformation-equivariance by either applying group convolutions in the network [9] or pre-warping [16, 20, 25, 43] the input image. But the above strategies need to exhaustively explore the possible transformation dimensions and degrees, as is illustrated in Fig. 1a, which is redundant in computation when coping with

*Corresponding author.

the homography transformation with a DOF of 6.

To cope with the above problem, homography-guided image warping, as shown in Fig. 1c, is adopted in our proposed recurrent homography estimation framework, dubbed RHWF. We note that homography-guided image warping has already been unconsciously employed in some of the previous cascading-based works [15, 22, 34]. However, the reason, effect, and technique of using homography-guided image warping, especially absorbing it properly in the recurrent framework, hasn't ever been investigated. Different from the previous works, our RHWF combines the homography-guided image warping with the recurrent trainable network, which significantly improves the accuracy without the cost of network parameters. Compared to the previous cascading-based SOTA method LocalTrans [34], RHWF reduces the mean average corner error (MACE) by about 70% on the MSCOCO dataset, while reducing the parameter cost of 86.5%.

On the other side, transformer architecture [8, 13] has demonstrated its superior ability in computer vision and image processing tasks. The transformer architecture has also been introduced in the homography estimation task as in [21, 34]. Following their pioneer exploration, we propose a transformer structure, named FocusFormer, that is pretty compatible with the homography-guided image warping and the recurrent framework. As illustrated in Fig. 1d, Fig. 1e, and Fig. 1f, unlike the attention mechanism in previous works that is pure global or local, FocusFormer employs the attention focusing mechanism. The scope of the attention mechanism shrinks along with the recurrence procedure, which captures the intra/inter correspondence information in a global→nonlocal→local[1] manner. We note that compared to the most widely adopted global attention mechanism, the attention-focusing mechanism can save computation costs while improving the homography estimation performance simultaneously.

We introduce the homography-guided image warping and FocusFormer into the recurrent homography estimation framework, named RHWF. The three parts, *i.e.,* recurrent estimation, homography-guided image warping, and the FocusFormer cooperate well, with each part facilitating the others. We evaluate RHWF on a variety of datasets including common RGB image data [24], cross-resolution data [34] and cross-modal data [6, 48], on which it outperforms all other competitors by a large gap. We show that though adopting the transformer, our RHWF reduces the parameter cost of 24.6% while achieving the accuracy gain of 38.1% (MSCOCO) and 34.1% (GoogleMap), compared to the previous SOTA method IHN [6]. In summary, our contributions are as follows: **(1)** We propose a novel Recurrent homography estimation framework using Homography-guided image Warping and FocusFormer, dubbed RHWF. RHWF ranks top on a variety of datasets, including the challenge scenes such as the cross-resolution and cross-modal ones. The recurrent estimation, homography-guided image warping, and FocusFormer facilitate the functionality of each other. **(2)** The reason, effect, and technique of using homography-guided image warping properly in the recurrent framework is first fully investigated. With the assistance of homography-guided image warping, the extracted features gradually converge into consistency, and hence boosting the homography estimation accuracy. **(3)** The FocusFormer is proposed to be the fundamental block of the recurrent homography estimation. The attention mechanism in FocusFormer works in a global→nonlocal→local manner, which significantly saves the computational costs while achieving a better performance.

[1] As in most of the works that refer to "nonlocal" [5], it denotes a relatively large neighborhood around a pixel.

2. Related Work

In this section we briefly review the most relevant works including deep homography estimation, transformation-equivariant network, and transformer in deep homography estimation. The readers are referred to literature such as [50] for the basic knowledge and traditional methods for homography estimation.

Deep Homography Estimation. DeTone *et al*. [12] first propose to estimate the homography deformation between the concatenated input image pair with a VGG-style network. Many works [15, 22, 46] inherit the network as a basic structure, which is either modified or cascaded by multiple times to boost the estimation accuracy. The cascading strategy indeed improves the performance, but is exceeded by the recurrent estimation methods [6, 7, 38, 48].

The recurrent homography estimation method mainly contains two types, including the IC-LK iterator based ones [7, 38, 48] and the deep iterator based ones [6]. CLKN [7] is the pioneer work introduced by Chang *et al*. to employ IC-LK to build an untrainable recurrent homography estimation framework. Zhao *et al*. [48] present a new loss function to directly enhance the similarity of the feature maps of CNNs, which further releases the potential of the IC-LK framework. However, the recurrent framework using IC-LK is unable to learn the implicit prior for the iterator through a large amount of image data. Cao *et al*. [6] propose IHN to make the whole network trainable, achieving a higher homography estimation accuracy.

Transformation-Equivariant Under Homography. It has been well analyzed in previous works [9] that standard convolution is equivariant to translations, but may fail to equivary with more general transformations. Unfortunately, homography has the transformation of rotation, scaling, shearing, aspect ratio, and perspective [43]. The corresponding features extracted by standard CNNs are inconsistent under homography deformation, which reduce the

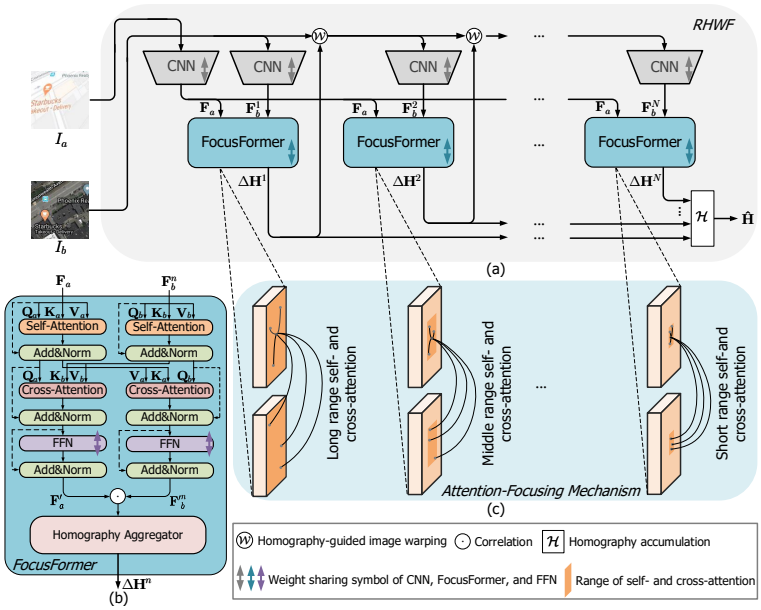


Figure 2. Architecture of Recurrent homography estimation using Homography-guided image Warping and FocusFormer, named RHWF. (a) Overall structure of RHWF. (b) Detailed structure of the proposed FocusFormer. (c) Detailed illustration of the attention-focusing mechanism.

estimation accuracy. Many efforts [9–11, 16, 25, 39] have been made to deal with this problem. Some methods design special network architectures to make CNNs equivariant to specific transformations [9–11, 39], among which the well-known group equivariant CNNs [9] produce rotation equivariant features using group convolution and subgroup pooling. However, the model becomes inefficient as it applies group convolutions directly on a large group [25]. The transformation is limited to rotation in this network. Other methods [16, 20, 25, 43] achieve transformation-equivariant by the predefined warping of the input image by different transformation dimensions and degrees. For example, GIFT [25] produces rotation and scale invariant features using warped images with predefined rotation angles and scale ratios. Warped convolution [20] achieves equivariance by warping the input image through a designed function. But the above two methods can only deal with the rotation/scaling of 2 DOF. When it comes to the homography transformation, the DOF of which increases from 2 to 6 (besides translation, which is equivariant in the standard convolution). The predefined warping will become burdensome and impracticable.

Transformer in Deep Homography Estimation. The performance of homography estimation networks can be boosted using the transformer. The transformer can be used either for feature enhancement or homography estimation. Shao *et al*. [34] adopt a multiscale transformer that works locally for the feature enhancement purpose. Hong *et al*. [21] employ a series of transformers to predict homography from the feature pyramids in a coarse-to-fine manner. The transformer shows powerful ability by significantly improving the homography estimation accuracy, and a transformer structure suitable for the recurrent homography framework is urgently needed.

3. Methodology

Fig. 2 illustrates the architecture of Recurrent homography estimation using Homography-guided image Warping and FocusFormer, named RHWF. The details of FocusFormer are shown in Fig. 2b. The homography-guided image warping and the attention-focusing mechanism in FocusFormer are arranged in an interleaved manner, which are appropriately absorbed into the recurrent framework. We note that the whole network is tied despite the number of iteration, which means the framework won't raise the network parameters. Let's denote the paired input images as I_a and I_b, and we aim to obtain the homography matrix $\mathbf{H}$ that relates them.

3.1. Backbone

We employ convolutional neural network (CNN) with residual connections as our backbone. As illustrated in Fig. 2a, we use CNNs with shared weights for 2 input images. The input images are processed with a convolutional block of kernel size 3×3 first. The produced feature maps are then continually processed by the basic blocks consisting of 2 residual layers, with each block accomplishing a 2×2 down-sampling in the spatial dimension. Different from the previous works [6, 34] that use the max-pooling layer to perform down-sampling, RHWF uses convolution with stride 2 to reduce the inference computation, which reduces the inference computational cost at 1 time to 1.43 GFLOPs

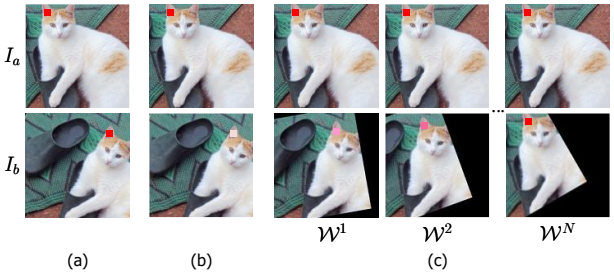


Figure 3. Illustration of the reason and effect of using homography-guided image warping. The red squares represent the feature extracted from the left-side ear of the cat in the images. The color of the squares exhibits the feature similarity, and the colors become consistent as the feature similarity grows. (a) The ideal homography-invariant situation. (b) The practical situation. (c) The effect of homography-guided image warping.

from 2.43 GFLOPs in [6]. Finally, the feature maps are projected by 1 linear convolutional layer of kernel size 1×1. RHWF only needs a dimension of 96 while achieving far better results than IHN [6] that needs 256. We stack 2 basic blocks to produce the feature maps of the spatial dimension $H/4 \times W/4$ and $H/2 \times W/2$, where H and W denote the height and width of the image. We note that, during the recurrent estimation process, the feature map of I_a is computed once but the one of I_b is computed recurrently due to the homography-guided image warping. This operation raises the computational cost of the network, but we will show the tremendous benefits of doing so.

3.2. Homography-Guided Image Warping

In previous recurrent methods [6, 7], the images of severe homography deformation are used directly to produce the feature maps for homography estimation. The defect that the standard convolution is lack of equivary under homography transformation is ignored. For the homography estimation task, we want the backbone network to extract homography equivariant feature maps, which can be formulated by

$$\phi(\pi(I; \mathbf{H})) = \pi(\phi(I); \mathbf{H}), \tag{1}$$

where π denotes the coordinate projection, $\mathbf{H}$ denotes the homography matrix that relates I_a and I_b, and ϕ denotes the backbone CNN. This ideal situation is illustrated in Fig. 3a. In the ideal situation, the homography deformation won't affect feature similarity, making 2 features absolutely match each other. Unfortunately, in practice, due to the lack of equivary under homography transformation, the similarity of corresponding features produced by the backbone CNN will be weakened as illustrated in Fig. 3b.

To cope with the above problem, many strategies have been proposed [9, 10, 20, 25] to either employ the group convolution or pre-warp the image by multiple times. However, both of the above methods bring considerable computational costs. We propose to interleave the homography estimation and image warping process, dubbed homography-guided image warping, in our RHWF. The whole process is designed under the inspiration of the alternate update of different variables in conventional optimization frameworks. For I_a and I_b related by homography $\mathbf{H}$, the interleaved image warping and homography estimation process can be expressed as

$$\begin{aligned} \mathbf{F}_b^n &= \phi(\mathcal{W}(I_b; \hat{\mathbf{H}}^n)), \\ \hat{\mathbf{H}}^{n+1} &= \psi(\mathbf{F}_a; \mathbf{F}_b^n), \end{aligned} \tag{2}$$

where $\mathcal{W}$ denote the homography-guided image warping, ψ denotes the FocusFormer[2] designed for homography estimation. The feature consistency between $\mathbf{F}_a$ and $\mathbf{F}_b$ together with the homography $\hat{\mathbf{H}}$ are optimized alternatively. We note that the homography-guided image warping takes the advantage of the recurrent framework of RHWF, and hence the redundant pre-designed warping or complex group convolution is not required. They are replaced by the warping guided by the present estimated homography. Meanwhile, the homography-guided image warping facilitates the estimation accuracy of the recurrent framework by relieving the feature inconsistency caused by deformation, as shown in Fig. 3c. We note that the backbone network recurrently extracts the feature map of $\mathbf{F}_b^n$ with tied weights, which won't bring additional parameter costs.

3.3. FocusFormer

The core architecture of RHWF is our FocusFormer that achieves recurrent residual homography estimation. FocusFormer mainly contains two parts that accomplish attention mechanism and homography estimation.

Attention. Following previous works [23, 34, 36, 40], we interleave the self-attention layer and the cross-attention layer, which is shown in Fig. 2b. By adopting both self-attention and cross-attention, FocusFormer can capture the intra/inter correspondence information of the input image pair.

As demonstrated in Fig. 2c, along with the network recurrence, the scope of attention mechanism shrinks in a global→nonlocal→local manner. This operation is deeply bounded with the homography-guided image warping process and the recurrent framework. At the first iteration, the deformation between the two images is large due to large homography deformation. Long range attention is needed to better capture the intra-inter correspondence. As the iteration continues, the deformation is continuously reduced, and hence the attention can focus on the local area to better improve the estimation accuracy. It is worth noting that the attention-focusing mechanism can reduce the computational cost to 3.18 GFLOPs from 7.14 GFLOPs, which

[2]In practice, the FocusFormer predicts the residual translation of 4 corner points of an image $\Delta\mathbf{T}^{n+1}$, and some simple computation will produce $\hat{\mathbf{H}}^{n+1}$ as will be detailed in Section 3.3.

works in the pure global manner.

Let us denote the coordinate index of the feature map as $\mathbf{x} = (u, v)$, the projected query, key, and value as $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$. The self attention-focusing mechanism can be formulated as

$$\text{self}(\mathbf{x}) = \text{softmax}\big(\frac{\mathbf{Q}_i(\mathcal{N}(\mathbf{x}))^\mathsf{T}\mathbf{K}_i(\mathcal{N}(\mathbf{x}))}{\sqrt{D}}\big)\mathbf{V}_i(\mathcal{N}(\mathbf{x})), \tag{3}$$

where $\mathcal{N}(\mathbf{x})$ represents the region around $\mathbf{x}$ that self-attention works within. If $\mathcal{N}(\mathbf{x})$ is set as the whole image, it would become the global self-attention, otherwise the non-local or local one of radius $R_\mathbf{A}$. The subscript "i" denotes that the query, key, and value are from the same image. The cross attention-focusing mechanism can be formulated as

$$\text{cross}(\mathbf{x}) = \text{softmax}\big(\frac{\mathbf{Q}_i(\mathcal{N}(\mathbf{x}))^\mathsf{T}\mathbf{K}_j(\mathcal{N}(\mathbf{x}))}{\sqrt{D}}\big)\mathbf{V}_j(\mathcal{N}(\mathbf{x})), \tag{4}$$

where subscripts "i" and "j" denote that the query is from one image, while the key and value are from another. As for other details, we only use 1 self-attention layer and 1 cross-attention layer. Considering that the whole attention block functions in the recurrent framework, the effectiveness of the attention layer would be fully motivated during the iterative inference process. The residual connecting, layer normalization, and feed-forward network are also adopted as in previous works [34, 40]. The multi-head attention is discarded as in [40] to improve efficiency.

Homography Estimation. After capturing the intra/inter correspondence of the input image pair, the correlation within local areas is then computed as

$$\mathbf{C}(\mathbf{x}, \mathbf{r}) = \text{ReLU}\big(\mathbf{F}'_a(\mathbf{x})^\mathsf{T}\mathbf{F}'^n_b(\mathbf{x}+\mathbf{r})\big), \quad \|\mathbf{r}\|_\infty \le R_\text{C} \tag{5}$$

where R_C controls the radius of each local area, making the correlation volume having the size of $H \times W \times (2R_\text{C}+1) \times (2R_\text{C}+1)$. This correlation can also be interpreted as another layer of cross attention [34] for the homography estimation, while the mapping differs as the desired output is the residual homography.

The architecture of the homography aggregator is similar to the previous correlation-based homography estimation works [6, 34]. The aggregator is composed of the basic blocks having 2 convolution layers and 1 max-pooling layer. The only difference is that the depth of the convolution layer of our RHWF (denoted as D) is set to be markedly lower than that of IHN [6] and LocalTrans [34], which also saves the network parameters. Like most previous works [6, 12, 34], we parameterize the residual homography in the form of the translation of the 4 corner points of an image, namely $\mathbf{T}$. At iteration n, The parameterization of $\hat{\mathbf{H}}^n$ using $\hat{\mathbf{T}}^n$ can be easily established by a least square problem as

$$\mathbf{A}^n\hat{\mathbf{h}}^n = \mathbf{b}^n, \tag{6}$$

where $\mathbf{b}^n$ is the coordinate of the projected 4 corner points, $\mathbf{A}^n$ is composed of the projected 4 corner points and the original 4 corner points, $\hat{\mathbf{h}}^n$ is the vectorized $\hat{\mathbf{H}}^n$. The original 4 corner points and the projected 4 corner points are related by $\hat{\mathbf{T}}^n$. During the recurrent process, the homography estimator produces the present residual translation $\Delta\mathbf{T}$, and the translation $\hat{\mathbf{T}}$ is updated by

$$\hat{\mathbf{T}}^{n+1} = \hat{\mathbf{T}}^n + \Delta\mathbf{T}^{n+1}, \tag{7}$$

which is equivalent to the homography update illustrated in Fig. 2a, which can be formulated as

$$\hat{\mathbf{H}}^{n+1} = \hat{\mathbf{H}}^n\Delta\mathbf{H}^{n+1}. \tag{8}$$

3.4. Multiscale Refinement

The performance of our proposed RHWF can be further improved by multiscale refinement. As illustrated in Section 3.1, the feature map of the spatial dimension $H/2 \times W/2$ can be adopted to construct an additional scale for a more accurate homography refinement. In the refinement scale, a FocusFormer with different weights is employed. The structures of the FocusFormer of the 2 levels are basically identical, except for the attention range of the additional scale, which further shrinks as the initial homography is given by the former scale. We will show that the accuracy of a single scale already outperforms most of the previous works except for the 2-scale IHN in [6].

3.5. Supervision

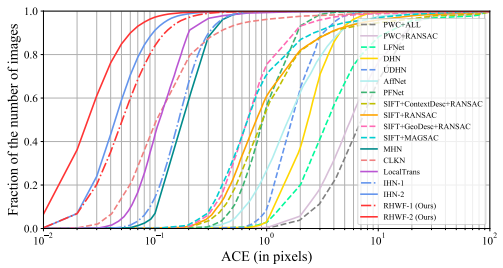
We use the L_1 loss between the ground-truth translation $\mathbf{T}_\text{gt}$ and the estimated one $\hat{\mathbf{T}}$. The loss at each iteration is weight summed as the final loss

$$L = \sum_{n=0}^{N-1}\gamma^{(N-n-1)}|\hat{\mathbf{T}}^n - \mathbf{T}_\text{gt}|, \tag{9}$$

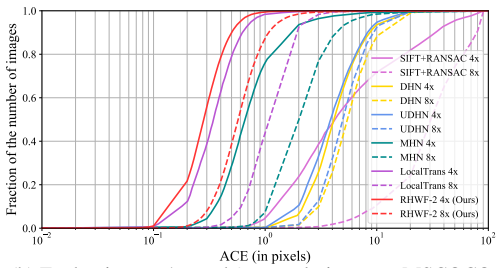
where N denotes the recurrent time within 1 scale, γ is less than 1 to produce a larger weight for the later estimations as in IHN [6]. If multiscale refinement is employed, the losses of the 2 scales are summed.

3.6. Implementation Details

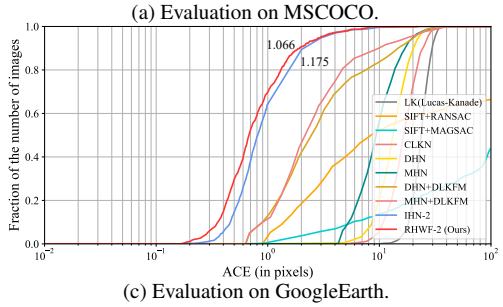
In our implementation, the recurrent time N within one scale is set to 6. We set the attention range in the FocusFormer as $[G, 4, 2, 1, 1, 1]$ for scale $H/4 \times W/4$ and $[2, 2, 1, 1, 1, 1]$ for scale $H/2 \times W/2$, where G denotes the global attention, and others the radius of the attention area R_A. The radius of correlation is set to $R_\text{C} = 8$ and $R_\text{C} = 4$ for the 2 scales. The depth of convolution layers is set to $D = 80$ and $D = 64$ for the 2 scales. The network is trained using AdamW [26] optimizer with the max learning rate of 0.0004 and the iteration of 120000.



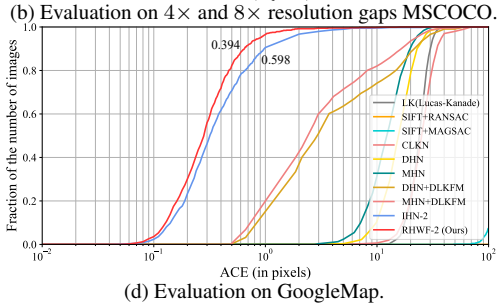
(a) Evaluation on MSCOCO.



(b) Evaluation on $4\times$ and $8\times$ resolution gaps MSCOCO.



(c) Evaluation on GoogleEarth.



(d) Evaluation on GoogleMap.

Figure 4. Homography estimation evaluation on MSCOCO, cross-resolution MSCOCO, GoogleEarth, and GoogleMap datasets. MSCOCO contains common RGB images. Cross-resolution MSCOCO includes image pairs of $4\times$ and $8\times$ resolution gaps. GoogleEarth and GoogleMap are cross-modal datasets. The suffixes "-1" and "-2" denote the scale of the network.

4. Experiments

4.1. Datasets and Experimental Setup

Datasets. We evaluate our RHWF on MSCOCO [24] following [6, 7, 12, 15, 22, 34, 48], $4\times$ and $8\times$ cross-resolution MSCOCO following [34], and cross-modal GoogleEarth together with GoogleMap following [6, 48]. MSCOCO is a large-scale real world RGB dataset, which is most widely used for homography estimation evaluation. The cross-resolution MSCOCO is employed in [34] to meet the homography estimation requirement in multiscale gigapixel photography [3, 47]. The cross-modal GoogleEarth and GoogleMap datasets are employed in [48] for the homography estimation requirement in the navigation scenario [18, 48]. We note that all the methods included for comparison together with our RHWF are trained and evaluated on the same training and test subset of each dataset for totally fair comparison.

Experimental setup. Similar to most of the previous works [6, 7, 12, 15, 22, 34, 48], the input images of size $[128 \times 128]$ are randomly perturbed in the corner points, with the perturbation range of $[-32, 32]$. The average corner error (ACE) is adopted for the homography accuracy evaluation following [6, 7, 12, 15, 22, 34, 48], which is lower when achieving higher accuracy.

4.2. Ablation

Ablation Study on MSCOCO Dataset. The ablation of the homography-guided image warping, the attention-focusing mechanism, and the multiscale refinement are shown in Table 1. Mean average corner error (MACE) and network parameters are also listed. All the ablations are conducted on the 1-scale RHWF unless otherwise specially mentioned. It is observed that the proposed homography-guided image warping and attention-focusing mechanism yields significant gain in accuracy. The global attention has a range of the whole image, and the local has an attention radius of 1. We can also find the multiscale refinement very effective by increasing the accuracy by 97.4%.

Table 1. Ablation study of RHWF.

Ablation part	Setting	MACE	Parameters
Warping	Feature warping	0.203(↑163.6%)	0.94 M
	Image warping	**0.077**	0.94 M
Attention	No	0.091(↑18.2%)	0.85 M
	Pure global	0.085(↑10.4%)	0.94 M
	Pure local	0.082(↑6.5%)	0.94 M
	Focus	**0.077**	0.94 M
Scale	**1 scale**	0.077(↑97.4%)	0.94 M
	2 scales	**0.039**	1.29 M

Ablation without recurrence. We conduct the ablation with the compared methods with no recurrence and the RHWF w/ (RHWF) and w/o (RHWF-NF) FocusFormer in Table 2. It is observed that the FocusFormer is effective.

Table 2. MACEs of RHWF and the compared structures without recurrent processing.

DHN	MHN	LocalTrans	IHN	RHWF	RHWF-NF
4.191	3.645	1.987	1.027	0.945	1.019

A Deeper Look at Warping. We further illustrate the average correlation value, which represents the feature similarity at each warping count in Fig. 5. The warping of feature or image is taken for comparison to evaluate the effectiveness of homography-guided image warping. It is observed that the image warping can significantly enhance the

feature similarity as the warping count increases, while the feature warping weakens it.

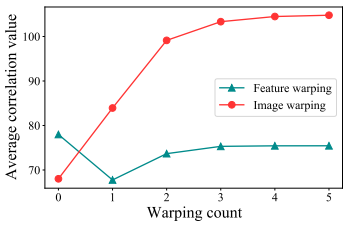


Figure 5. The average correlation value at each warping count of feature warping and image warping.

4.3. MSCOCO Evaluation

We evaluate RHWF together with other homography estimation methods including IHN [6] (previous SOTA method), LocalTrans [34], MHN [22], UDHN [46], DHN [12], CLKN [7], AffNet [30], LFNet [33], PFNet [42], PWC [35], SIFT+ContextDesc+RANSAC [28], SIFT+GeoDesc+RANSAC [29], SIFT+MAGSAC [2], and SIFT+RANSAC [27]. The comparison is illustrated in Fig. 4a. Following previous works [6,7,22,34,48], we plot the fraction of the number of images w.r.t. corresponding ACEs of a dataset. It can be observed that our 1-scale RHWF already outperforms other competitors except for 2-scale IHN, while 2-scale RHWF outperforms all others with a large gap. We note that 2-scale RHWF reduces previous SOTA MACE by 38.1%. We also plot the result of 1-scale IHN, which is significantly exceeded by 1-scale RHWF. Please kindly note that CLKN performs IC-LK iteration with 4 scales, LocalTrans and MHN cascade 3 scales of networks to improve the estimation accuracy, but our 1-scale RHWF already markedly outperforms them.

We then plot the MACEs at each iteration for IHN and our RHWF that adopt the recurrent framework in Fig. 6. It is observed that RHWF exceeds IHN at the 3rd iteration, while a more iterations won't make IHN outperform the RHWF with only 3 iterations. What's more, RHWF continues to reduce MACEs obviously with more iterations.

4.4. Cross-Resolution MSCOCO Evaluation

We conduct the evaluation on $4\times$ and $8\times$ cross-resolution MSCOCO following [34]. We compare our RHWF with LocalTrans [34] (previous SOTA method), MHN [22], UDHN [46], DHN [12], and SIFT+RANSAC [27]. The resolution gap dramatically increases the difficulty of homography estimation, while it is vital for multi-scale gigapixel photography [3, 34]. The results of $4\times$ and $8\times$ resolution gap MSCOCO are plotted in Fig. 4b. It is observed that RHWF ranks top in both scenarios and is significantly more robust under the $8\times$ resolution gap.

We further demonstrate the self- and cross-attention map of RHWF at each iteration in Fig. 7. It is observed that at the

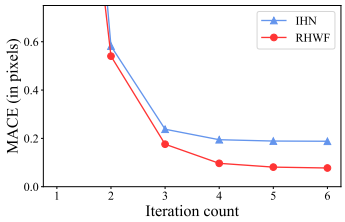


Figure 6. The MACE comparison of IHN and our RHWF at each iteration count.

1st iteration, the global attention in I_a and $I_b(4\downarrow)$ successfully captures the correspondence information. However, in $I_b(8\downarrow)$, the global attention becomes ambiguous as the resolution gap grows. Fortunately, as the recurrence continues, the homography-guided image warping corrects the deformation and the attention-focusing mechanism shrinks the attention range, which gradually clarifies the attention targets.

4.5. Cross-Modal Datasets Evaluation

The cross-modal data further raises the challenge for homography estimation, while it can be employed for GPS-denied navigation [18]. GoogleEarth contains image pairs across different seasons and GoogleMap the image pairs cropped from satellite images and their corresponding map images. We include IHN [6] (previous SOTA method), the original LK [1], SIFT+RANSAC [27], SIFT+MAGSAC [2], CLKN [7], DHN [12], MHN [22], DHN+DLKFM [48], and MHN+DLKFM [48] for comparison. The results of both datasets are separately demonstrated in Fig. 4c and Fig. 4d. The MACEs of RHWF and IHN are also illustrated for a better comparison. RHWF outperforms IHN by 34.1% on GoogleMap[3].

We further illustrate the homography estimation results of the above-mentioned methods except for CLKN, DHN+DLKFM, and LK, as they are outperformed by MHN+DLKFM in [48]. It can be observed in Fig. 8 that under the severe large deformation and modality gap, SIFT+RANSAC and SIFT+MAGSAC fail as in the experiment of [6]. MHN+DLKFM produces unstable results as the LK iterator is of theoretical drawback [32]. DHN, MHN, and IHN are not as accurate as RHWF, as they neither consider the feature inconsistency caused by homography deformation nor introduce an attention mechanism to capture the intra-inter correspondence information. On the contrary, our RHWF can produce promising results under large deformation and modality gaps.

[3]We notice that the performance of RHWF on GoogleEarth is not that conspicuous compared to other datasets but still outperforms other methods. This is because GoogleEarth only contains about 8k image pairs of fixed homography deformations, which limits the ability of the model by disabling more deformation augmentation.

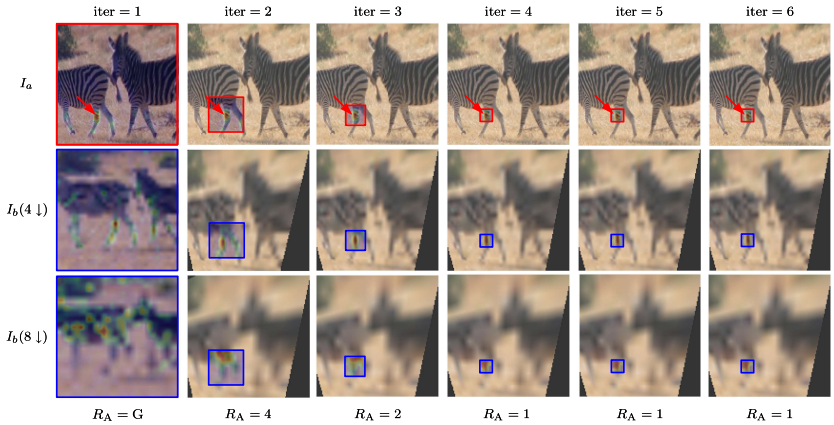


Figure 7. The self- and cross-attention map of RHWF at each iteration. The 1st row: the image I_a of the standard resolution with the self-attention map. The 2nd and 3rd row: the image I_b with $4\times$ ($I_b(4\downarrow)$) and $8\times$ ($I_b(8\downarrow)$) downsampling and cross-attention maps. The red arrows denote the query point of attention, and the red and blue boxes separately highlight the self- and cross-attention maps.

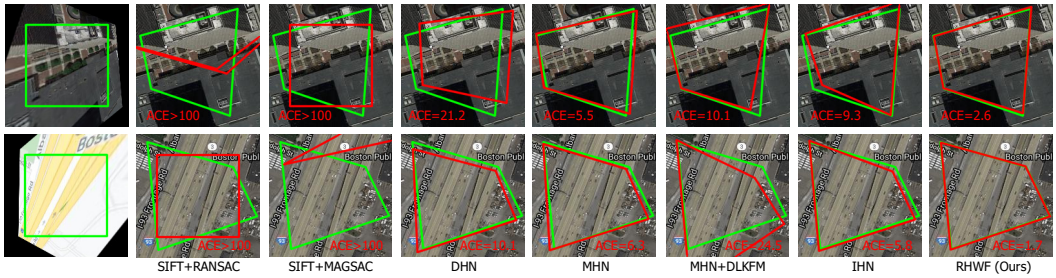


Figure 8. Homography estimation results of methods including IHN [6], SIFT+RANSAC [27], SIFT+MAGSAC [2], CLKN [7], DHN [12], MHN [22], MHN+DLKFM [48], and our RHWF.

Table 3. Parameter comparison.

RHWF	IHN	LocalTrans	DHN	MHN	UDHN	DLKFM
1.29 M	1.71 M	9.56 M	34.19 M	2.57 M	21.29 M	19.24 M

4.6. Parameter and FLOPs Comparison

We conduct the parameter and FLOPs comparison in Table 3. It is observed that compared to previous methods, our RHWF owns the least parameter cost, which is reduced by 86.5% and 24.6 % compared with previous SOTA works LocalTrans [34] and IHN [6].

In Table 4, we compare the FLOPs of models and MACEs on MSCOCO with the previous SOTA recurrent methods including IHN, its improved version IHN-mov, and DLKFM. Compared to IHN-mov and DLKFM, the computational costs spent on the homography-guided warping in RHWF are much more effective. We also take the RHWF of the recurrent time 3 (fewer FLOPs), namely RHWF-3, which outperforms IHN-mov and DLKFM.

Table 4. FLOPs of models with MACEs on MSCOCO.

	RHWF	RHWF-3	IHN	IHN-mov	DLKFM
FLOPs	16.96 G	9.64 G	8.34 G	20.32 G	110.51 G
MACE	0.077	0.176	0.191	0.177	0.550

4.7. Failure Case

Fig. 9 shows 2 failure cases, the left one is caused by the error accumulation, and the right the imperfect initial estimations.

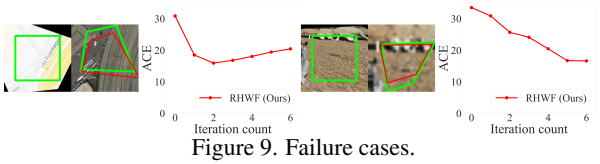


Figure 9. Failure cases.

5. Conclusions

We have proposed a novel recurrent homography estimation framework, named RHWF. RHWF absorbs the homography-guided image warping and the FocusFormer, which facilitate the homography estimation by enhancing the feature consistency and capturing the intra/inter corresponding information in a global→nonlocal→local manner, into the recurrent framework. Experimentally, RHWF outperforms previous methods by a large gap with significantly fewer parameters. The computation cost is raised by the homography-guided image warping and attention operation, which is the limitation of our proposed framework.

Acknowledgement

This work was supported in part by the "Pioneer" and "Leading Goose" R & D Program of Zhejiang under grant 2023C03136 and in part by the Ten Thousand Talents Program of Zhejiang Province under grant 2020R52003. We also thank the generous help from Jun Ma, Zhejiang University and Tianyu Guo, Peking University.

References

[1] Simon Baker and Iain Matthews. Lucas-Kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 56(3):221–255, 2004. 1, 7

[2] Daniel Barath, Jiri Matas, and Jana Noskova. MAGSAC: marginalizing sample consensus. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10197–10205, 2019. 7, 8

[3] David J Brady, Michael E Gehm, Ronald A Stack, Daniel L Marks, David S Kittle, Dathon R Golish, EM Vera, and Steven D Feller. Multiscale gigapixel photography. *Nature*, 486(7403):386–389, 2012. 1, 6, 7

[4] Matthew Brown and David G Lowe. Automatic panoramic image stitching using invariant features. *International journal of computer vision*, 74(1):59–73, 2007. 1

[5] Thomas Brox, Oliver Kleinschmidt, and Daniel Cremers. Efficient nonlocal means for denoising of textural patterns. *IEEE Transactions on Image Processing*, 17(7):1083–1092, 2008. 2

[6] Si-Yuan Cao, Jianxin Hu, Zehua Sheng, and Hui-Liang Shen. Iterative deep homography estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1879–1888, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[7] Che-Han Chang, Chun-Nan Chou, and Edward Y Chang. CLKN: Cascaded lucas-kanade networks for image alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2213–2221, 2017. 1, 2, 4, 6, 7, 8

[8] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021. 2

[9] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, pages 2990–2999. PMLR, 2016. 1, 2, 3, 4

[10] Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical CNNs. In *International Conference on Learning Representations*, 2018. 3, 4

[11] Taco S Cohen and Max Welling. Steerable CNNs. *arXiv preprint arXiv:1612.08498*, 2016. 3

[12] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Deep image homography estimation. *arXiv preprint arXiv:1606.03798*, 2016. 1, 2, 5, 6, 7, 8

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[14] Jakob Engel, Thomas Schöps, and Daniel Cremers. LSD-SLAM: Large-scale direct monocular SLAM. In *Proceedings of the European Conference on Computer Vision*, pages 834–849. Springer, 2014. 1

[15] Farzan Erlik Nowruzi, Robert Laganiere, and Nathalie Japkowicz. Homography estimation from image pairs with hierarchical convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 913–920, 2017. 1, 2, 6

[16] Yujie Fu, Pengju Zhang, Bingxi Liu, Zheng Rong, and Yihong Wu. Learning to reduce scale differences for large-scale invariant image matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 1, 3

[17] Junhong Gao, Seon Joo Kim, and Michael S Brown. Constructing image panoramas using dual-homography warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 49–56. IEEE, 2011. 1

[18] Hunter Goforth and Simon Lucey. GPS-denied UAV localization using pre-existing satellite imagery. In *2019 International Conference on Robotics and Automation*, pages 2974–2980. IEEE, 2019. 1, 6, 7

[19] Heng Guo, Shuaicheng Liu, Tong He, Shuyuan Zhu, Bing Zeng, and Moncef Gabbouj. Joint video stitching and stabilization from moving cameras. *IEEE Transactions on Image Processing*, 25(11):5491–5503, 2016. 1

[20] Joao F Henriques and Andrea Vedaldi. Warped convolutions: Efficient invariance to spatial transformations. In *International Conference on Machine Learning*, pages 1461–1469. PMLR, 2017. 1, 3, 4

[21] Mingbo Hong, Yuhang Lu, Nianjin Ye, Chunyu Lin, Qijun Zhao, and Shuaicheng Liu. Unsupervised homography estimation with coplanarity-aware GAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17663–17672, 2022. 1, 2, 3

[22] Hoang Le, Feng Liu, Shu Zhang, and Aseem Agarwala. Deep homography estimation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7652–7661, 2020. 1, 2, 6, 7, 8

[23] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6197–6206, 2021. 4

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014. 2, 6

[25] Yuan Liu, Zehong Shen, Zhixuan Lin, Sida Peng, Hujun Bao, and Xiaowei Zhou. Gift: Learning transformation-invariant dense visual descriptors via group CNNs. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 3, 4

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[27] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 7, 8

[28] Zixin Luo, Tianwei Shen, Lei Zhou, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. Contextdesc: Local descriptor augmentation with cross-modality context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2527–2536, 2019. 7

[29] Zixin Luo, Tianwei Shen, Lei Zhou, Siyu Zhu, Runze Zhang, Yao Yao, Tian Fang, and Long Quan. Geodesc: Learning local descriptors by integrating geometry constraints. In *Proceedings of the European Conference on Computer Vision*, pages 168–183, 2018. 7

[30] Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Repeatability is not enough: Learning affine regions via discriminability. In *Proceedings of the European Conference on Computer Vision*, pages 284–300, 2018. 7

[31] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015. 1

[32] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006. 1, 7

[33] Yuki Ono, Eduard Trulls, Pascal Fua, and Kwang Moo Yi. LF-Net: Learning local features from images. *arXiv preprint arXiv:1805.09662*, 2018. 7

[34] Ruizhi Shao, Gaochang Wu, Yuemei Zhou, Ying Fu, Lu Fang, and Yebin Liu. Localtrans: A multiscale local transformer network for cross-resolution homography estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14890–14899, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[35] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018. 7

[36] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 4

[37] Richard Szeliski. Image alignment and stitching: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 2(1):1–104, 2006. 1

[38] Chaoyang Wang, Hamed Kiani Galoogahi, Chen-Hsuan Lin, and Simon Lucey. Deep-LK for efficient adaptive object tracking. In *2018 IEEE International Conference on Robotics and Automation*, pages 627–634. IEEE, 2018. 2

[39] Maurice Weiler and Gabriele Cesa. General e (2)-equivariant steerable CNNs. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[40] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. GMFlow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8121–8130, 2022. 4, 5

[41] Jiacheng Ying, Hui-Liang Shen, and Si-Yuan Cao. Unaligned hyperspectral image fusion via registration and interpolation modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 2021. 1

[42] Rui Zeng, Simon Denman, Sridha Sridharan, and Clinton Fookes. Rethinking planar homography estimation using perspective fields. In *Asian Conference on Computer Vision*, pages 571–586. Springer, 2018. 7

[43] Xinrui Zhan, Yang Li, Wenyu Liu, and Jianke Zhu. Warped convolution networks for homography estimation. *arXiv preprint arXiv:2206.11657*, 2022. 1, 2, 3

[44] Xinrui Zhan, Yueran Liu, Jianke Zhu, and Yang Li. Homography decomposition networks for planar object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3234–3242, 2022. 1

[45] Haoxian Zhang and Yonggen Ling. Hvc-net: Unifying homography, visibility, and confidence learning for planar object tracking. In *Proceedings of the European Conference on Computer Vision*, pages 701–718. Springer, 2022. 1

[46] Jirong Zhang, Chuan Wang, Shuaicheng Liu, Lanpeng Jia, Nianjin Ye, Jue Wang, Ji Zhou, and Jian Sun. Content-aware unsupervised deep homography estimation. In *Proceedings of the European Conference on Computer Vision*, pages 653–669. Springer, 2020. 2, 7

[47] Jianing Zhang, Tianyi Zhu, Anke Zhang, Xiaoyun Yuan, Zihan Wang, Sebastian Beetschen, Lan Xu, Xing Lin, Qionghai Dai, and Lu Fang. Multiscale-vr: Multiscale gigapixel 3d panoramic videography for virtual reality. In *2020 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2020. 6

[48] Yiming Zhao, Xinming Huang, and Ziming Zhang. Deep Lucas-Kanade homography for multimodal image alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15950–15959, 2021. 1, 2, 6, 7, 8

[49] Yuan Zhou, Anand Rangarajan, and Paul D Gader. An integrated approach to registration and fusion of hyperspectral and multispectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 58(5):3020–3033, 2019. 1

[50] Barbara Zitova and Jan Flusser. Image registration methods: A survey. *Image and Vision Computing*, 21(11):977–1000, 2003. 2