Semi-supervised Deep Large-Baseline Homography Estimation with Progressive Equivalence Constraint

Hai Jiang[1,3*], **Haipeng Li**[2,3*], **Yuhang Lu**[4], **Songchen Han**[1,†], **Shuaicheng Liu**[2,3,†]
[1]Sichuan University
[2]University of Electronic Science and Technology of China
[3]Megvii Technology
[4]University of South Carolina
{jianghai1@stu., hansongchen@}scu.edu.cn, {lihaipeng@std.,liushuaicheng@}uestc.edu.cn
{jianghai,lihaipeng,liushuaicheng}@megvii.com, yuhang@email.sc.edu

Abstract

Homography estimation is erroneous in the case of large-baseline due to the low image overlay and limited receptive field. To address it, we propose a progressive estimation strategy by converting large-baseline homography into multiple intermediate ones, cumulatively multiplying these intermediate items can reconstruct the initial homography. Meanwhile, a semi-supervised homography identity loss, which consists of two components: a supervised objective and an unsupervised objective, is introduced. The first supervised loss is acting to optimize intermediate homographies, while the second unsupervised one helps to estimate a large-baseline homography without photometric losses. To validate our method, we propose a large-scale dataset that covers regular and challenging scenes. Experiments show that our method achieves state-of-the-art performance in large-baseline scenes while keeping competitive performance in small-baseline scenes. Code and dataset are available at https://github.com/megvii-research/LBHomo.

Introduction

Homography estimation is a basic task in computer vision that has been widely used for a wide range of high-level vision tasks, such as image/video stitching (Guo et al. 2016), video stabilization (Liu et al. 2013), SLAM (Du et al. 2020), and HDR reconstruction (Liu et al. 2021b, 2022). Traditional methods typically use feature detection and matching algorithms (Lowe 2004; Rublee et al. 2011), and subsequently solve direct linear transform ($\mathcal{DLT}$) (Hartley and Zisserman 2003) with outlier suppression to obtain a homography matrix. However, these methods are highly dependent on the extracted feature matches and may fail in scenes that lack sufficient high-quality feature points. On the other hand, learning-based methods take a pair of source and target images as input and directly output the corresponding homography matrix. They do not rely on matched key points and thus are more robust than traditional methods. The learning-based methods can be divided into two categories: supervised methods (DeTone, Malisiewicz, and Rabinovich 2016) and unsupervised methods (Nguyen et al. 2018). The supervised ones use synthetic image pairs for training due to the lack of sufficient real-world image pairs with GT labels, lacking realistic scene parallax causes unsatisfactory generalization ability. Unsupervised methods adopt label-free training strategies that aim to minimize the photometric distance between the warped source images and target images, being better generalized to various scenes. With the assistance of photometric losses (Zhang et al. 2020; Ye et al. 2021), the unsupervised methods perform well in small-baseline scenes where the non-overlap rate between two images is less than 10%. However, in large-baseline scenes where the non-overlap rate is between 20% and 50%, the warped source image contains a number of out-of-boundary pixels due to the large appearance and viewpoint changes, causing it hard to minimize the photometric distance. As shown in Fig. 1, BasesHomo (Ye et al. 2021) can successfully align two images with large overlap but fails in

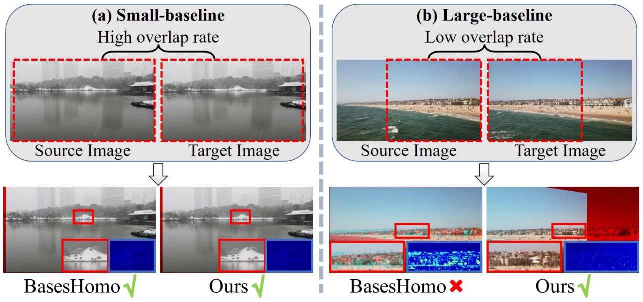

Figure 1: Comparison of our method with an unsupervised learning-based method, i.e., BasesHomo. The first row shows two pairs of consecutive frames in different baseline cases, we highlight the overlap region in red boxes. In the second row, we superimpose the warped source image and the target image, where misaligned pixels are visualized as colored ghosts. Local details are enlarged in red and blue boxes, where the latter is an error heatmap, the darker the better. Our method can handle both situations, while BasesHomo fails in the large-baseline scene.

[*] These authors contributed equally.
[†] Corresponding authors

small overlap, while our method is capable to handle both cases. To address large-baseline cases, (Nie et al. 2021) proposed an ablation-based strategy, which forces the two images to contain the same size of valid areas to ignore out-of-boundary pixels. However, this strategy fails where dynamic objects exist in images. From the above, we find it is non-trivial to estimate the homography of two images with a large-baseline. To this end, we propose a progressive homography estimation strategy to address the large-baseline challenge. Specifically, we convert the large-baseline problem into multiple intermediate phases by inserting several images along with pre-defined homographies into the source and target image, as shown in Fig. 2(a). The homography between the source image and the target image can be obtained by cumulatively multiplying these intermediate homographies. To achieve this, we design a homography identity loss to optimize our network in a semi-supervised way by considering the advantages and weaknesses of supervised and unsupervised learning. Our loss function consists of two parts: a supervised objective and an unsupervised objective. Firstly, given the pre-defined homographies, the supervised one is utilized to optimize intermediate homographies of inserted images. Secondly, the unsupervised loss helps to estimate a large-baseline homography without photometric loss. More specifically, given the target image, our network predicts two homographies, the former being the last intermediate image to it and the later being the source towards it. Our unsupervised loss minimizes the error between the homography of the source and target image and the accumulative multiplication result of all intermediate items, i.e. the homographies of inserted images and the homography of the last intermediate and target image, as shown in Fig. 2(b).

With our progressive estimation strategy and semi-supervised loss, the network can focus on homography optimization while avoiding the problems caused by photometric losses. Additionally, we introduce a large-scale real-world image pairs dataset for large-baseline homography estimation considering there lacks a dedicated dataset for such a task, which contains 5 categories of scenes as well as human-labeled GT point correspondences for quantitative evaluation. Extensive experiments demonstrate that our method outperforms the state-of-the-art supervised methods and unsupervised methods both quantitatively and qualitatively. Moreover, our method is also applicable to small-baseline scenarios. To summarize, our main contributions are threefold:

- We propose a progressive estimation strategy to address large-baseline homography estimation by transforming the large-baseline into several intermediate ones.
- We propose a semi-supervised homography identity loss that enforces the network to focus on optimizing the homography.
- We introduce a large-scale dataset containing various scenes for large-baseline homography learning and the human-labeled evaluation set is also included. Experimental results demonstrated that our method achieves state-of-the-art performance.

Related Work

Traditional Homography Estimation

Traditional homography estimation methods usually combine classic or learning-based feature extraction and matching algorithms such as SIFT (Lowe 2004), ORB (Rublee et al. 2011), BEBLID (Suárez et al. 2020), SuperPoint (DeTone, Malisiewicz, and Rabinovich 2018), SOSNet (Tian et al. 2019), SuperGlue (Sarlin et al. 2020), LoFTR (Sun et al. 2021), and subsequently solve direct linear transform with outliner suppression such as RANSAC (Fischler and Bolles 1981), MAGSAC (Barath, Matas, and Noskova 2019). However, feature-based methods usually crash in challenging scenes where sufficient feature matches cannot be obtained. In addition, some methods can also solve a homography directly by using the Lucas-Kanade algorithm (Baker and Matthews 2004) or calculating the sum of squared differences (SSD) between two images without extracted feature matches. A randomly initialized homography is optimized in this way iteratively.

Deep Homography Estimation

Following the development of learning-based image alignment methods, such as optical flow (Li, Luo, and Liu 2021; Luo et al. 2021; Han et al. 2022) and dense correspondence (Truong et al. 2021a; Liu et al. 2021a), a deep homography estimation network was first proposed by (DeTone, Malisiewicz, and Rabinovich 2016). Deep homography estimation methods can be divided into two categories: supervised and unsupervised. The former ones (Shao et al. 2021; Cao et al. 2022) utilize the generated image pairs with ground-truth labels to train their models, but their generalization ability is limited due to the lack of realistic scene parallax in synthetic images. Unsupervised methods (Nguyen et al. 2018; Kharismawati et al. 2020; Ye et al. 2021) optimize their models with real-world image pairs by minimizing the photometric distance from the source image warped by the estimated homography to the target image. To be more robust, some methods (Zhang et al. 2020; Hong et al. 2022) introduce efficient masks to replace classic outlier rejection methods to remove undesired regions or focus on the dominant plane. However, most of the previous methods are proposed to estimate the homography of image pairs with a small-baseline, large-baseline homography estimation, a field with broader applications, has long been ignored.

Deep Image Alignment

Recently, some image stitching methods (Nie et al. 2021, 2020) use an individual homography estimation network for coarse alignment and optimize the pre-aligned images by reconstruction networks to achieve better stitching results in large-baseline scenes. However, their networks are not specially designed for such tasks, leading to unsatisfactory results. In addition, some geometric matching methods (Nie et al. 2022a; Truong, Danelljan, and Timofte 2020; Truong et al. 2021b) can also be applied to solve the large-baseline image alignment. But their alignment is mainly realized by mesh flow or dense flow, which contains the local motion information of the images, while the homography matrix

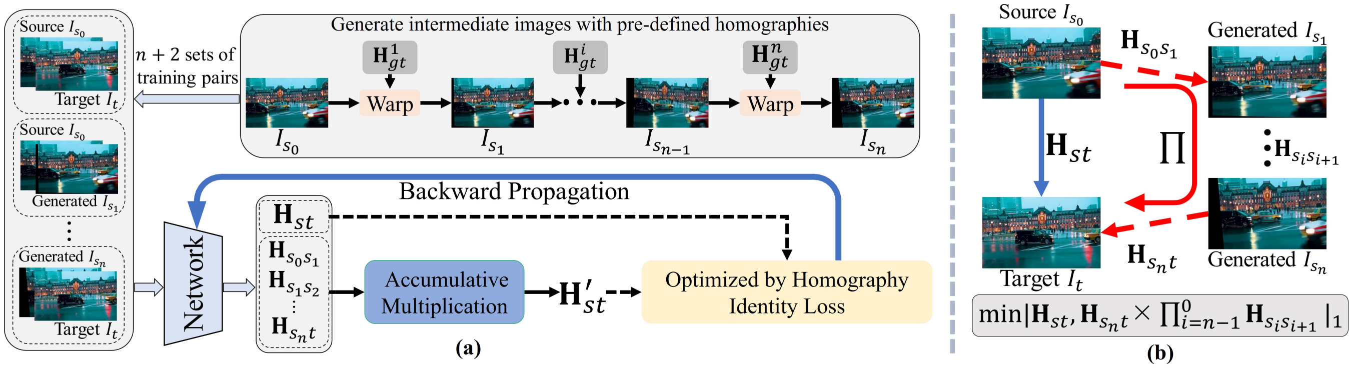

Figure 2: The left column (a) illustrates our proposed progressive estimation strategy, and right column (b) illustrates our proposed unsupervised objective function. $\prod$ denotes the accumulative multiplication operation.

only represents the global motion. In contrast, we propose a progressive estimation strategy and a semi-supervised consistency constraint without photometric loss to address the large-baseline homography estimation.

Method

Overview

In this section, we introduce a progressive strategy to address large-baseline homography estimation. Specifically, we transform the large-baseline into several intermediate stages by inserting some intermediate images, i.e., $I_{s_i} \in \mathbb{R}^{H\times W\times 3}$ ($i \in [0, n]$), into the source image $I_s \in \mathbb{R}^{H\times W\times 3}$ and target image $I_t \in \mathbb{R}^{H\times W\times 3}$. We generate the intermediate ones by a set of pre-defined homographies, i.e., $I_{s_{i+1}} = \mathcal{W}_{i+1}(I_{s_i}, \mathbf{H}_{gt}^{i+1})$, where $\mathcal{W}_{i+1}$ represents the warp operation by the $\mathbf{H}_{gt}^{i+1}$ and I_{s_0} is the initial I_s. By randomly sampling the pre-defined homographies with non-identity matrices, it is effective to avoid the degenerate solutions of accumulative multiplication results of intermediate homographies (Truong et al. 2021b). In addition, we ensure that the non-overlap rate of the two inserted images is smaller than that of the source and target image.

After generating n intermediate images, we get $n + 2$ sets of image pairs, i.e., $(I_{s_i}, I_{s_{i+1}})$, (I_{s_n}, I_t), and (I_s, I_t). Our goal is to train a neural network f_θ, with parameters θ, that predicts homography matrix $\mathbf{H}_{s_is_{i+1}} = f_\theta(I_{s_i}, I_{s_{i+1}})$, $\mathbf{H}_{s_nt} = f_\theta(I_{s_n}, I_t)$, $\mathbf{H}_{st} = f_\theta(I_s, I_t)$ relating I_{s_i} to $I_{s_{i+1}}$, I_{s_n} to I_t, and I_s to I_t respectively. Multiplying all of the intermediate homographies $\mathbf{H}_{s_is_{i+1}}$ and the $\mathbf{H}_{s_nt}$ should be equal to the $\mathbf{H}_{st}$. With this equivalence constraint, we can enforce the network to optimize the homographies themselves. To achieve this, we propose a semi-supervised homography identity loss to train our network, which is described in the following section.

Homography Identity Loss

The point-wise correspondence between each set of image pairs can be the mapping of the corresponding homography matrix. Let us denote $\mathbf{X}_{s_i}$ and $\mathbf{X}_t$ as the meshgrid coordinate sets of I_{s_i} and I_t, respectively. The point-wise correspondence of $\mathbf{X}_{s_i}$ and $\mathbf{X}_{s_{i+1}}$ is associated by the $\mathbf{H}_{s_is_{i+1}}$ using $\mathbf{X}_{s_{i+1}} = \mathbf{H}_{s_is_{i+1}}\mathbf{X}_{s_i}$. Likewise, the correspondence of $\mathbf{X}_{s_n}$ and $\mathbf{X}_t$ is related by the $\mathbf{H}_{s_nt}$. Accordingly, the coordinate correspondence between the source and target image can be expressed as $\mathbf{X}_t = \mathbf{H}_{st}\mathbf{X}_s = (\mathbf{H}_{s_nt} \times \mathbf{H}_{s_{n-1}s_n} \times \cdots \times \mathbf{H}_{s_0s_1})\mathbf{X}_s$, where $\times$ denotes the cross-product operation. Therefore, the homography of the source image and the target image $\mathbf{H}_{st}$ can be obtained by multiplying the $\mathbf{H}_{s_nt}$ and $\mathbf{H}_{s_is_{i+1}}$. The corresponding identity equation can be expressed as

$$\prod_{i=n-1}^{0} \mathbf{H}_{s_is_{i+1}} = \mathbf{H}_{s_nt}^{-1} \times \mathbf{H}_{st}. \tag{1}$$

In essence, our optimization goal is not to minimize the distance between the warped source image and the target image, but to minimize the error between estimated homographies based on Eq.(1) in an unsupervised manner. The unsupervised objective function ($\mathcal{L}_{unsup}$) is formulated as

$$\mathcal{L}_{unsup} = |\mathbf{H}_{s_nt}^{-1} \times \mathbf{H}_{st} - \prod_{i=n-1}^{0} \mathbf{H}_{s_is_{i+1}}|_1, \tag{2}$$

where $|\cdot|_1$ denotes the L1 norm. Since the intermediate images are generated through the pre-defined homographies, so we can estimate the homographies of intermediate images in a supervised way. The supervised objective function ($\mathcal{L}_{sup}$) is formulated as

$$\mathcal{L}_{sup} = \sum_{i=0}^{n-1} |\mathbf{H}_{s_is_{i+1}} - \mathbf{H}_{gt}^{i+1}|_1, \tag{3}$$

and the $\mathbf{H}_{s_is_{i+1}}$ in Eq.(2) can be replaced with the $\mathbf{H}_{gt}^{i+1}$. However, due to the cancellation effect amid the estimated homography terms, artlessly replacing $\mathbf{H}_{s_is_{i+1}}$ with $\mathbf{H}_{gt}^{i+1}$ may obtain degeneration solutions, e.g., $\mathbf{H}_{st} = I_{3\times3}$. To avoid this situation, we rewrite Eq.(2) as

$$\mathcal{L}_{unsup} = \sum_{i=0}^{n-1} \lambda_i |(\mathbf{H}_{s_{i+1}t}^{-1} \times \mathbf{H}_{st}) - \prod_{j=i}^{0} \mathbf{H}_{gt}^{j+1}|_1. \tag{4}$$

Our final semi-supervised homography identity loss $\mathcal{L}_{HIL}$ combines the $\mathcal{L}_{sup}$ and $\mathcal{L}_{unsup}$ as $\mathcal{L}_{HIL} = \mathcal{L}_{unsup} + \lambda_w\mathcal{L}_{sup}$, where the λ_w is a weighting factor, we eliminate this hyper-parameter by automatically balancing the weights over each training batch as $\lambda_w = \mathcal{L}_{unsup}/\mathcal{L}_{sup}$.

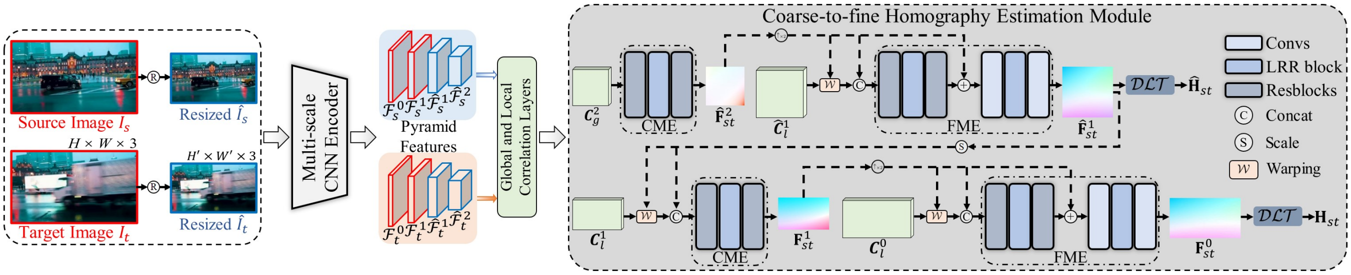

Figure 3: The overall pipeline of our proposed network. Firstly, we resize the original images into lower resolution ones and use a multi-scale CNN encoder to obtain the pyramid features. The correlation layers are used to improve the utilization of feature information and expand the receptive field. Finally, we use a coarse-to-fine homography estimation module to obtain homography flows and solve $\mathcal{DLT}$ to get the corresponding homography matrices.

Multi-Scale Homography Estimation Network

There are two challenges for the large-baseline homography estimation: 1) the overlap rate between two images is low, and 2) the receptive field of CNN-based models is limited (Nie et al. 2022b). To overcome these problems, we design a multi-scale homography estimation network, as shown in Fig. 3, that combines a multi-scale CNN encoder and correlation layers to leverage feature information and expand the receptive field. Subsequently, the global and local correlation maps are fed into a coarse-to-fine homography estimation module to obtain the final results.

Multi-scale CNN Encoder. Given a pair of images I_s and I_t, we use a multi-scale CNN encoder consists of four cascaded residual blocks and max-pooling layers to extract multi-scale features as $\mathcal{F}^k \in \mathbb{R}^{\frac{H}{2^{2+k}} \times \frac{W}{2^{2+k}} \times d^k}$, $k \in [0, 2]$. Features at the top pyramid layer have the lowest resolution, representing the most global information, which are subsequently used to generate the global correlation map. The size of the global correlation map is determined by the size of the input features, which requires a significant computation in the case of high-resolution images. Therefore, we resize the I_s and I_t to $\hat{I}_s$ and $\hat{I}_t$ with fixed lower resolution $H^{'} \times W^{'}$. The $\hat{I}_s$ and $\hat{I}_t$ are fed into the encoder to produce features as $\hat{\mathcal{F}}^k$. As shown in Fig. 3, we select $\mathcal{F}^0$, $\mathcal{F}^1$, $\hat{\mathcal{F}}^1$, and $\hat{\mathcal{F}}^2$ to form a four-layer feature pyramid. A set of homographies can be estimated from the pyramid layers, which are transmitted in a coarse-to-fine manner.

Feature Correlation. Previous unsupervised methods (Zhang et al. 2020; Ye et al. 2021; Hong et al. 2022) estimate homography without using correlation layers. However, we find that the correlation is effective for feature matching. Specifically, the $\hat{\mathcal{F}}^2$ contains more global information than others, we use the global correlation layer to represents the pairwise similarity between spatial positions in the source feature $\hat{\mathcal{F}}_s^2$ and target feature $\hat{\mathcal{F}}_t^2$ as

$$\mathbf{C}_g^2(\mathbf{x}_s, \mathbf{x}_t) = \hat{\mathcal{F}}_s^2(\mathbf{x}_s)^\top \hat{\mathcal{F}}_t^2(\mathbf{x}_t), \tag{5}$$

where $\mathbf{x}_s$ and $\mathbf{x}_t$ denote the coordinate position of the source and target feature, and $\mathbf{C}_g^2$ is the global correlation map of the source and target feature. The result is a 4D tensor, we reshape it to a 3D tensor of size $\frac{H^{'}}{16} \times \frac{W^{'}}{16} \times \frac{H^{'} \times W^{'}}{256}$.

For the rest feature maps, i.e., $\mathcal{F}^0$, $\mathcal{F}^1$, $\hat{\mathcal{F}}^1$, we apply the local correlation layer proposed in (Truong et al. 2020) to evaluate the feature similarity between two feature maps, denoted as $\mathbf{C}_l^0$, $\mathbf{C}_l^1$, and $\hat{\mathbf{C}}_l^1$. The search region R is set to constrain the search space and result in local correlation maps with the size of $\mathbf{C}_l$ is $\frac{H^{(')}}{2^{2+k}} \times \frac{W^{(')}}{2^{2+k}} \times (2R+1)^2$.

Coarse-to-Fine Homography Estimation Module. Given feature correlations, we adopt two coarse motion estimators (CME) and two fine motion estimators (FME) to extract the global and local motion from the correlations. As discussed above, the relative motion between I_s and I_t can be mapped by a homography. Therefore, We compute the homography flow (Li, Luo, and Liu 2021) as $\mathbf{F}_{st} = \mathbf{X}_t - \mathbf{X}_s$ to facilitate the learning of motion information and subsequently solve the $\mathcal{DLT}$ to obtain a unique homography. The homography flow of $\hat{I}_s$ and $\hat{I}_t$ is expressed as $\hat{\mathbf{F}}_{st}$. Specifically, the $\mathbf{C}_g^2$ is fed into the CME to estimate the coarse homography flow $\hat{\mathbf{F}}_{st}^2$, and the fine $\hat{\mathbf{F}}_{st}^1$ is obtained by combining $\hat{\mathbf{F}}_{st}^2$ and $\hat{\mathbf{C}}_l^1$ as input through the FME. Likewise, the homography flow $\mathbf{F}_{st}^1$ and $\mathbf{F}_{st}^0$ can be generated with the corresponding features and correlations via motion estimators. Moreover, the LRR blocks (Ye et al. 2021) are inserted at FME to reject motion outliers implicitly, detailed architectures of the CME and FME are illustrated in Fig 3. Finally, we convert the flows into the corresponding homographies by solving the $\mathcal{DLT}$.

Network Training

As mentioned in (DeTone, Malisiewicz, and Rabinovich 2016), it is non-trivial to directly estimate a homography matrix. Therefore, we use the homography flow as the supervision object during the training stage. Besides, we resize the initial images to fixed-resolution to improve the applicability to high resolution images, our supervised and unsupervised objectives are accordingly converted into

$$\begin{aligned} \mathcal{L}_{unsup} = \sum_{i=0}^{n-1} \lambda_i (|\mathbf{F}_{st} - \mathbf{F}_{s_{i+1}t} - \sum_{j=i}^{0} \mathbf{F}_{gt}^{j+1}|_1 \\ + |\hat{\mathbf{F}}_{st} - \hat{\mathbf{F}}_{s_{i+1}t} - \sum_{j=i}^{0} \hat{\mathbf{F}}_{gt}^{j+1}|_1), \end{aligned} \tag{6}$$

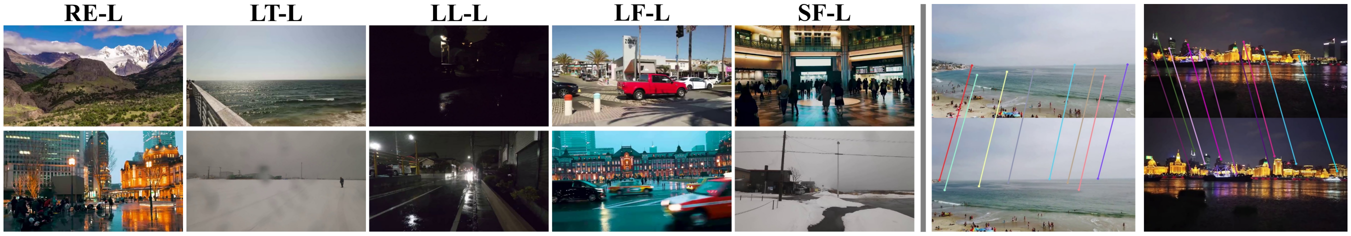

Figure 4: A glace of our dataset. The dataset contains five categories as shown in the first five columns. The rightmost two columns show several examples of human labeled point correspondences for quantitative evaluation.

$$\mathcal{L}_{sup} = \sum_{i=0}^{n-1} |\mathbf{F}_{s_i s_{i+1}} - \mathbf{F}_{gt}^{i+1}|_1 + \sum_{i=0}^{n-1} |\hat{\mathbf{F}}_{s_i s_{i+1}} - \hat{\mathbf{F}}_{gt}^{i+1}|_1. \quad (7)$$

The most relative representations of our homography flow are the 8 bases flow (Ye et al. 2021) and optical flow. While the former performs well in small-baseline scenes, it crashes in large-baselines. Compared to optical flow, with the assistance of our supervised objective and LRR blocks, our homography flows tend to represent the global motion between two images. Importantly, the homography flows are only used to facilitate training, and $\mathcal{DLT}$ will be used to transform the flows into homography matrices.

Experiments

Dataset. We introduce a large-scale dataset for large-baseline homography estimation considering there lacks a dedicated dataset for this task. Our dataset contains 5 categories, including regular (RE-L), low-texture (LT-L), low-light (LL-L), small-foregrounds (SF-L), and large-foregrounds (LF-L) scenes. We select image pairs from real-world scenes and ensure the average non-overlap rate between the source and target images is from 20% to 50%. Our dataset contains ≈78k image pairs in totally, and 1.8k image pairs are randomly chosen from all categories as the evaluation data. For each evaluation image pair, we manually labeled more than 6 uniform distributed matching points for quantitative comparisons. Some examples of our dataset are illustrated in Fig. 4.

Implementation Details. In the training stage, we randomly crop patches of size 320×480 near the center of the initial images as input, and the resolution of resized images is set to $(H^{'} \times W^{'}) = (256 \times 256)$. The number of inserted images is set to $n = 2$, and the non-overlap rate of the two intermediate images is less than 20%. We empirically set the λ_i in Eq.(7) to 10^{-i}. Our network is implemented with PyTorch, and the training is performed on four NVIDIA RTX 2080Ti GPUs. The Adam optimizer (Kingma and Ba 2015) is adopted with an initial learning rate of 5×10^{-4} for model optimization, and it decays by a factor of 0.8 after every epoch. The batch size is set to 16.

Comparison with Existing Methods

Quantitative Comparison. We report the quantitative results of all comparison methods on our large-baseline evaluation set in Table 1, where rows 3-8 are traditional feature-based methods, rows 9-14 are learned feature-based methods, rows 15-18 are deep homography estimation methods, and rows 19-22 are deep image alignment methods which can be applied to large-baseline scenes. $I_{3\times3}$ in the first row refers to the identity transformation, of which the errors reflect the original distance between point pairs. And we have retrained all deep learning-based methods, except the Supervised (DeTone, Malisiewicz, and Rabinovich 2016), on our training set represented by *. As shown in Table 1, our method achieves state-of-the-art performance in all categories of the large-baseline dataset. In the regular (RE-L) scenes, our method and SIFT+RANSAC achieve the best results, because the feature-based methods can obtain sufficient high-quality matching points in regular scenes and thus perform well. But feature-based methods fail in other challenging scenes, especially in low-light (LL-L) and low-texture (LT-L), while our method does not rely on feature detection and correspondence matching, being more robust than traditional methods in these scenarios. For example, our method reduces the error on LT-L by 66.4% compared to SIFT+MAGSAC. Moreover, image alignment methods perform better than feature-based methods in some challenging scenarios, our method still produces lower errors, e.g., our method reduces the error on LL-L by 53.0% compared to UNSUPDIS. The small-foreground (SF-L) and large-foreground (LF-L) scenes contain dynamic objects, affecting the estimation of homography. Compared with other deep learning-based methods using photometric losses for optimization, our method outperforms them in LF-L and SF-L benefiting from our homography identity loss.

Additionally, we also conduct experiments on the small-baseline dataset (Zhang et al. 2020), which also contains 5 categories RE-S, LT-S, LL-S, LF-S, and SF-S. Considering the small relative motion between two images in small-baseline scenes, we set $n = 1$ and reduce the non-overlap rate of the source and inserted image. As reported in Table 2, our method outperforms the existing four deep learning-based methods with the error reduced from 0.50 to 0.44.

Qualitative Comparison. We compare the qualitative results of our method and other methods on our large-baseline dataset. Fig. 5 shows the visualization results of our method and four other methods in large-baseline scenes. Our method shows superiority in LT-L and LL-L scenes, where feature-based methods all fail due to insufficient key points extracted, as shown in Fig. 5(b) and Fig. 5(c). The other two deep image alignment methods can not align these images as well as ours. More specifically, they try to align the island

1)		RE-L	LT-L	LL-L	LF-L	SF-L	Avg-L
2)	$I_{3\times3}$	94.60 (+15428.0%)	106.48 (+550.9%)	99.42 (+1036.2%)	43.57 (+407.8%)	67.27 (+718.4%)	82.27 (+698.7%)
3)	SIFT+RAN	1.66 (+0.0%)	26.47 (+61.8%)	20.37 (+132.8%)	9.05 (+5.5%)	8.22 (+0.0%)	13.15 (+27.7%)
4)	SIFT+MAG	1.74 (+13.3%)	16.36 (+0.0%)	28.74 (+228.5%)	12.39 (+44.4%)	12.19 (+48.3%)	14.29 (+38.7%)
5)	ORB+RAN	5.74 (+677.3%)	71.91 (+339.6%)	25.19 (+187.9%)	19.08 (+122.4%)	44.36 (+439.7%)	33.26 (+222.9%)
6)	ORB+MAG	6.68 (+833.3%)	74.21 (+353.6%)	26.52 (+203.1%)	19.85 (+131.4%)	44.82 (+445.3%)	34.42 (+234.2%)
7)	BEBLID+RAN	19.68 (2991.3%)	90.50 (+453.2%)	52.32 (+497.9%)	30.76 (+258.5%)	61.32 (+646.0%)	50.92 (+394.4%)
8)	BEBLID+MAG	21.26 (+3253.6%)	90.96 (+456.0%)	53.87 (+515.7%)	31.99 (+272.8%)	61.36 (+646.5%)	51.89 (+403.8%)
9)	SOSNet+RAN	1.86 (+33.2%)	29.76 (+81.9%)	18.92 (+116.2%)	13.59 (+58.4%)	8.96 (+9.0%)	14.61 (+41.8%)
10)	SOSNet+MAG	2.02 (+59.8%)	35.42 (+116.5%)	19.18 (+119.2%)	19.02 (+121.7%)	11.95 (+45.4%)	17.52 (+70.1%)
11)	SuperPoint+RAN	1.74 (+13.3%)	32.29 (+97.4%)	12.97 (+48.2%)	11.83 (+37.9%)	14.90 (+81.3%)	14.75 (+43.2%)
12)	SuperPoint+MAG	1.86 (+33.2%)	35.19 (+115.1%)	13.34 (+52.5%)	12.90 (+50.4%)	14.28 (+73.7%)	15.51 (+50.6%)
13)	LoFTR+RAN	1.73 (+11.6%)	16.85 (+3.0%)	25.96 (+196.7%)	8.58 (+0.0%)	10.52 (+28.0%)	12.73 (+23.6%)
14)	LoFTR+MAG	1.78 (+19.9%)	18.27 (+11.7%)	26.31 (+200.7%)	9.89 (+15.3%)	11.74 (+42.8%)	13.60 (+32.0%)
15)	Supervised	94.59 (+15426.0%)	106.46 (+550.7%)	99.41 (+1036.2%)	43.40 (+405.9%)	67.13 (+716.7%)	82.20 (+698.1%)
16)	Unsupervised*	94.13 (+15349.8%)	105.94 (+547.6%)	98.76 (+1028.7%)	43.35 (+405.2%)	67.01 (+715.3%)	81.84 (+694.6%)
17)	CAHomo*	92.11 (+15015.0%)	99.91 (+510.7%)	91.27 (+943.1%)	35.80 (+317.2%)	58.72 (+614.3%)	75.56 (+633.6%)
18)	BasesHomo*	77.07 (+12518.7%)	95.76 (+485.3%)	81.67 (+833.3%)	32.99 (+284.5%)	49.15 (+498.0%)	67.33 (+553.7%)
19)	UNSUPDIS	3.49 (+303.8%)	34.00 (+107.8%)	8.75 (+0.0%)	12.29 (+43.2%)	10.93 (+33.0%)	13.89 (+34.9%)
20)	UNSUPDIS*	3.40 (+288.8%)	34.88 (+113.2%)	9.26 (+5.8%)	11.05 (+28.8%)	9.46 (+15.1%)	13.61 (+32.1%)
21)	DAMG	2.55 (+147.7%)	19.54 (+19.4%)	9.10 (+4.0%)	11.74 (+36.8%)	10.03 (+22.0%)	10.59 (+2.8%)
22)	DAMG*	2.17 (+84.7%)	20.12 (+23.0%)	8.78 (+0.3%)	10.82 (+26.1%)	9.62 (+17.0%)	10.30 (+0.0%)
23)	Ours	**1.66 (+0.0%)**	**5.49 (-66.4%)**	**4.11 (-53.0%)**	**7.57 (-11.8%)**	**6.95 (-15.5%)**	**5.16 (-49.9%)**

Table 1: The point matching errors (PME) of our method and all comparison methods on our large-baseline dataset. The best results are highlighted in bold, the second best results are underlined. The percentages in the parentheses indicate the relative change in comparison to the second best results. RAN and MAG denote RANSAC and MAGSAC, respectively.

1)		RE-S	LT-S	LL-S	LF-S	SF-S	Avg-S
2)	$I_{3\times3}$	7.75	7.65	7.21	3.39	7.53	6.70
3)	Supervised	1.51	4.48	2.76	3.00	2.62	2.87
4)	Unsupervised	0.79	2.45	1.48	1.10	1.11	1.39
5)	CAHomo	0.73	1.01	1.03	0.70	0.92	0.88
6)	BasesHomo	**0.29**	**0.54**	0.65	0.41	0.61	0.50
7)	Ours	**0.29**	0.64	**0.45**	**0.39**	**0.45**	**0.44**

Table 2: The point matching errors (PME) of our method and deep learning-based comparison methods on the small-baseline dataset. The best results are highlighted in bold.

1)	Modification	RE-L	LT-L	LL-L	LF-L	SF-L	Avg-L
2)	$n = 1$	1.91	6.49	4.71	7.11	7.26	5.50
3)	$n = 3$	1.64	9.77	4.05	8.70	7.92	6.42
4)	$\mathcal{L}_{PL}$	59.41	75.47	53.56	30.91	48.29	53.53
5)	$\mathcal{L}_{ABPL}$	5.81	18.56	15.47	10.62	10.69	12.23
6)	Only $\mathcal{L}_{sup}$	1.78	10.41	5.25	10.87	9.18	7.50
7)	Only $\mathcal{L}_{unsup}$	2.50	8.63	5.84	6.78	7.07	6.16
8)	Corner offsets	4.73	13.21	9.31	15.19	15.09	11.51
9)	8 bases flow	7.14	13.04	13.18	12.50	14.56	12.09
10)	Default	1.66	5.49	4.11	7.57	6.95	5.16

Table 3: Results of ablation studies, please refer to the text for more details.

in Fig. 5(b) since it has more texture than the surrounding area, while only our method successfully aligns the scene. In Fig. 5(d) and Fig. 5(e), UNSUPDIS and DAMG even perform worse than some feature-based methods because their optimization strategy is based on photo losses and thus cannot obtain satisfactory results in scenes with dynamic objects. Our method avoids the drawbacks of the photometric losses and thus generates more accurate results. Please refer to the supplementary material for more qualitative results.

Robustness Evaluation. Furthermore, we evaluate the robustness of all comparison methods on the large-baseline scenes by setting thresholds to calculate the proportion of inlier predictions. Specifically, points with errors lower than the threshold are considered inliers. As such, we set the threshold from 1.0 to 50.0. Our method significantly outperforms other methods from threshold 10.0 to 50.0. With a threshold of 10, our inlier proportion is 7.7% higher than the second best (89.1% vs. 81.4%), and our method does not produce any points with an error greater than 50.

Ablation Studies

We conduct extensive ablation studies to verify the effectiveness of our proposed components, and the results are reported in Table 3.

Progressive Estimation Strategy. In this experiment, we choose to vary the number of inserted intermediate images to verify the effectiveness of our progressive estimation strategy, as shown in rows 2 and 3. With only one image inserted, i.e., $n = 1$, the average error of our method is 5.50, which already surpasses other comparison methods in Table 1, but can still be reduced. After inserting three images, i.e., $n = 3$, the average error increased by 1.26 compared to $n = 2$. This is because the more images are inserted, the more accurate intermediate results should be interpolated, otherwise, the cumulative error increases significantly (Liu et al. 2009,

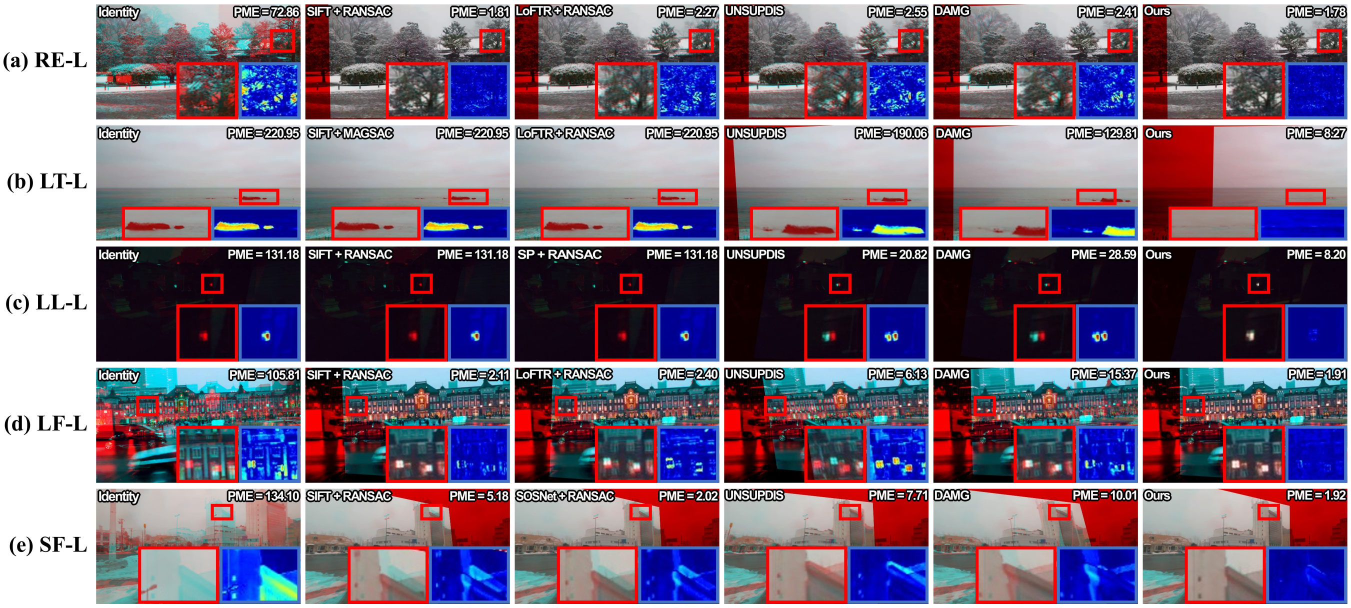

Figure 5: Qualitative results of our method and other competitive methods on the large-baseline dataset. The images are generated by superimposing the warped source images on the target image. Error-prone regions are highlighted with red boxes, and the blue boxes show the content difference between two images in the error-prone regions. Best viewed with zooming in.

2013). Even though the number of inserted images varies, our strategy still achieves superior results compared to other methods, and the optimal results can be obtained when the number of inserted images is moderate.

Homography Identity Loss. To verify the effectiveness of our homography identity loss, we first compare our loss function with the photometric losses used in (Ye et al. 2021) and (Nie et al. 2021), as shown in rows 4 and 5. We can see that optimizing with photometric loss (Ye et al. 2021), i.e., $\mathcal{L}_{PL}$ leads to failure in all scenarios, which is consistent with what we have discussed in the Introduction. The ablation-based photometric loss (Nie et al. 2021), i.e., $\mathcal{L}_{ABPL}$, avoids the effect of out-of-boundary pixels, it is not able to handle scenes with dynamic objects. Our proposed loss performs better than both of them in all scenes, which demonstrates the usefulness of our loss function. In addition, our homography identity loss is a semi-supervised loss, to verify the effectiveness of semi-supervised learning, we remove the $\mathcal{L}_{sup}$ and $\mathcal{L}_{unsup}$ respectively for optimization. As shown in rows 6 and 7, the average error of training with $\mathcal{L}_{sup}$ only is higher than that of training with $\mathcal{L}_{unsup}$ only (7.50 vs. 6.16), which is due to the fact that our supervised objective is constructed based on the synthetic data and is therefore not ideal in terms of generalizability. But solely using $\mathcal{L}_{unsup}$ is worse than solely using $\mathcal{L}_{sup}$ in RE-L and LL-L scenarios. By combining the advantages of supervised and unsupervised learning, our semi-supervised loss can achieve better results.

Homography Flow. In the training stage, we adopt the homography flow to facilitate network training. From another perspective, our homography flow is equivalent to a form of dense offsets, similar to the commonly used corner offsets (Nguyen et al. 2018; Zhang et al. 2020). However, our homography flow contains more motion information and achieves better results, as shown in row 8. Another similar form is the 8 bases flow (Ye et al. 2021), it performs well in the small-baseline scenes but crashes in the large-baseline, as shown in row 9.

Limitations

Although our method achieves state-of-the-art performance in large-baseline scenes compared with the existing methods, it still has its limitation of being applied to scenes with multiple planes where a homography theoretically cannot perform alignment well. We will leave the solution for the multiple planes alignment as future work.

Conclusion

In this work, we have presented a new deep framework for large-baseline homography estimation. We note that it is non-trivial to directly estimate a large-baseline homography and thus propose a progressive estimation strategy to convert it into several intermediate phases. The homography of two images can be obtained by cumulatively multiplying these intermediate ones. Meanwhile, we propose a semi-supervised homography identity loss to enforce the network focus on optimizing the homography itself, avoiding the problems of photometric losses. Moreover, we present a large-scale dataset for large-baseline homography estimation, which consists of five categories of scenes. Extensive experiments and ablation studies prove the effectiveness of our newly proposed components and demonstrate the superiority of our method over the existing methods.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (NSFC) under grant No.61872067 and in part by the Natural Science Foundation of Sichuan Province (NSFSC) under grant No.2023NSFSC0462.

References

Baker, S.; and Matthews, I. 2004. Lucas-kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 56(3): 221–255.

Barath, D.; Matas, J.; and Noskova, J. 2019. MAGSAC: marginalizing sample consensus. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10197–10205.

Cao, S.-Y.; Hu, J.; Sheng, Z.; and Shen, H.-L. 2022. Iterative Deep Homography Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1879–1888.

DeTone, D.; Malisiewicz, T.; and Rabinovich, A. 2016. Deep image homography estimation. *arXiv preprint arXiv:1606.03798*.

DeTone, D.; Malisiewicz, T.; and Rabinovich, A. 2018. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 224–236.

Du, S.; Guo, H.; Chen, Y.; Lin, Y.; Meng, X.; Wen, L.; and Wang, F.-Y. 2020. GPO: Global Plane Optimization for Fast and Accurate Monocular SLAM Initialization. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 6254–6260.

Fischler, M. A.; and Bolles, R. C. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6): 381–395.

Guo, H.; Liu, S.; He, T.; Zhu, S.; Zeng, B.; and Gabbouj, M. 2016. Joint video stitching and stabilization from moving cameras. *IEEE Transactions on Image Processing*, 25(11): 5491–5503.

Han, Y.; Luo, K.; Luo, A.; Liu, J.; Fan, H.; Luo, G.; and Liu, S. 2022. RealFlow: EM-Based Realistic Optical Flow Dataset Generation from Videos. In *European Conference on Computer Vision*, 288–305.

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

Hong, M.; Lu, Y.; Ye, N.; Lin, C.; Zhao, Q.; and Liu, S. 2022. Unsupervised Homography Estimation with Coplanarity-Aware GAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17663–17672.

Kharismawati, D. E.; Akbarpour, H. A.; Aktar, R.; Bunyak, F.; Palaniappan, K.; and Kazic, T. 2020. CorNet: Unsupervised Deep Homography Estimation for Agricultural Aerial Imagery. In *European Conference on Computer Vision*, 400–417.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representation*.

Li, H.; Luo, K.; and Liu, S. 2021. GyroFlow: Gyroscope-Guided Unsupervised Optical Flow Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12869–12878.

Liu, F.; Gleicher, M.; Jin, H.; and Agarwala, A. 2009. Content-preserving warps for 3D video stabilization. *ACM Transactions on Graphics*, 28(3): 1–9.

Liu, S.; Li, H.; Wang, Z.; Wang, J.; Zhu, S.; and Zeng, B. 2021a. Deepois: Gyroscope-guided deep optical image stabilizer compensation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5): 2856–2867.

Liu, S.; Yuan, L.; Tan, P.; and Sun, J. 2013. Bundled camera paths for video stabilization. *ACM Transactions on Graphics*, 32(4): 1–10.

Liu, Z.; Lin, W.; Li, X.; Rao, Q.; Jiang, T.; Han, M.; Fan, H.; Sun, J.; and Liu, S. 2021b. ADNet: Attention-guided deformable convolutional network for high dynamic range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 463–470.

Liu, Z.; Wang, Y.; Zeng, B.; and Liu, S. 2022. Ghost-free High Dynamic Range Imaging with Context-aware Transformer. In *European Conference on Computer Vision*, 344–360.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2): 91–110.

Luo, K.; Wang, C.; Liu, S.; Fan, H.; Wang, J.; and Sun, J. 2021. UPFlow: Upsampling Pyramid for Unsupervised Optical Flow Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1045–1054.

Nguyen, T.; Chen, S. W.; Shivakumar, S. S.; Taylor, C. J.; and Kumar, V. 2018. Unsupervised deep homography: A fast and robust homography estimation model. *IEEE Robotics and Automation Letters*, 3(3): 2346–2353.

Nie, L.; Lin, C.; Liao, K.; Liu, M.; and Zhao, Y. 2020. A view-free image stitching network based on global homography. *Journal of Visual Communication and Image Representation*, 73: 102950.

Nie, L.; Lin, C.; Liao, K.; Liu, S.; and Zhao, Y. 2021. Unsupervised deep image stitching: Reconstructing stitched features to images. *IEEE Transactions on Image Processing*, 30: 6184–6197.

Nie, L.; Lin, C.; Liao, K.; Liu, S.; and Zhao, Y. 2022a. Depth-Aware Multi-Grid Deep Homography Estimation With Contextual Correlation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7): 4460–4472.

Nie, L.; Lin, C.; Liao, K.; and Zhao, Y. 2022b. Learning edge-preserved image stitching from multi-scale deep homography. *Neurocomputing*, 491: 533–543.

Rublee, E.; Rabaud, V.; Konolige, K.; and Bradski, G. 2011. ORB: An efficient alternative to SIFT or SURF. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2564–2571.

Sarlin, P.-E.; DeTone, D.; Malisiewicz, T.; and Rabinovich, A. 2020. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4938–4947.

Shao, R.; Wu, G.; Zhou, Y.; Fu, Y.; Fang, L.; and Liu, Y. 2021. Localtrans: A multiscale local transformer network for cross-resolution homography estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14890–14899.

Suárez, I.; Sfeir, G.; Buenaposada, J. M.; and Baumela, L. 2020. BEBLID: Boosted efficient binary local image descriptor. *Pattern Recognition Letters*, 133: 366–372.

Sun, J.; Shen, Z.; Wang, Y.; Bao, H.; and Zhou, X. 2021. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8922–8931.

Tian, Y.; Yu, X.; Fan, B.; Wu, F.; Heijnen, H.; and Balntas, V. 2019. Sosnet: Second order similarity regularization for local descriptor learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11016–11025.

Truong, P.; Danelljan, M.; Gool, L. V.; and Timofte, R. 2020. GOCor: Bringing globally optimized correspondence volumes into your neural network. *Advances in Neural Information Processing Systems*, 33: 14278–14290.

Truong, P.; Danelljan, M.; and Timofte, R. 2020. GLU-Net: Global-Local Universal Network for Dense Flow and Correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Truong, P.; Danelljan, M.; Van Gool, L.; and Timofte, R. 2021a. Learning Accurate Dense Correspondences and When To Trust Them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5714–5724.

Truong, P.; Danelljan, M.; Yu, F.; and Van Gool, L. 2021b. Warp consistency for unsupervised learning of dense correspondences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10346–10356.

Ye, N.; Wang, C.; Fan, H.; and Liu, S. 2021. Motion basis learning for unsupervised deep homography estimation with subspace projection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13117–13125.

Zhang, J.; Wang, C.; Liu, S.; Jia, L.; Ye, N.; Wang, J.; Zhou, J.; and Sun, J. 2020. Content-aware unsupervised deep homography estimation. In *European Conference on Computer Vision*, 653–669.