Supervised Homography Learning with Realistic Dataset Generation

Hai Jiang[1,2,*], Haipeng Li[3,2,*], Songchen Han[1,†], Haoqiang Fan[2], Bing Zeng[3], Shuaicheng Liu[3,2,†]
[1]Sichuan University [2]Megvii Technology
[3]University of Electronic Science and Technology of China
{jianghai1@stu., hansongchen@}scu.edu.cn, {jianghai,lihaipeng,fhq,liushuaicheng}@megvii.com
{lihaipeng@std.,eezeng@,liushuaicheng@}uestc.edu.cn

Abstract

In this paper, we propose an iterative framework, which consists of two phases: a generation phase and a training phase, to generate realistic training data and yield a supervised homography network. In the generation phase, given an unlabeled image pair, we utilize the pre-estimated dominant plane masks and homography of the pair, along with another sampled homography that serves as ground truth to generate a new labeled training pair with realistic motion. In the training phase, the generated data is used to train the supervised homography network, in which the training data is refined via a content consistency module and a quality assessment module. Once an iteration is finished, the trained network is used in the next data generation phase to update the pre-estimated homography. Through such an iterative strategy, the quality of the dataset and the performance of the network can be gradually and simultaneously improved. Experimental results show that our method achieves state-of-the-art performance and existing supervised methods can be also improved based on the generated dataset. Code and dataset are available at https://github.com/JianghaiSCU/RealSH.

1. Introduction

Homography estimation is a fundamental task in computer vision that has been widely used for high-dynamic range imaging [40, 24, 25], image stitching [32, 31], and video stabilization [23, 43]. Traditional methods typically use feature extraction and matching methods [26, 3] with outlier suppression [9, 2] to obtain feature matches of two images and subsequently solve direct linear transformation (DLT) [11] to obtain the homography matrix. However, these methods highly rely on the extracted matching keypoints and may crash in challenging scenes that lack sufficient high-quality feature matches. With the rise of deep learning, such problems have been partially solved, learning-based methods [22, 20, 13] take a pair of images as input and directly output the corresponding homography, thus are more robust than traditional methods due to their keypoint-free estimation strategy. The learning-based methods can be divided into two categories: supervised and unsupervised. At the moment, benefiting from the label-free characteristic, unsupervised methods can be trained on enormous amounts of real-world data, delivering better performance and generalization capability than supervised ones. We find that the lack of qualified training data is one of the main barriers to the development of supervised methods. Previously supervised dataset generation strategy [6] synthesizes an image pair by using pre-defined ground truth (GT) homography to warp a single image, such strategy considers the whole image as a single plane, neglecting par-

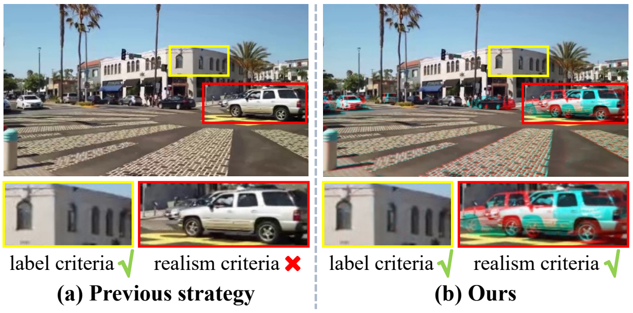


Figure 1. The first row shows the target images generated by the previous strategy [6] and our proposed method, visualized by superimposing the source image warped by GT homography on the generated target image, where the misaligned pixels are represented as colored ghosts. As highlighted in yellow boxes, the dominant plane is fully aligned by the GT homography, proving that both our method and the previous strategy can satisfy the label criteria. In contrast, only our method can maintain realistic motion between the two images thus satisfying the realism criteria, as shown in the red boxes, the car is a moving object that cannot be aligned by a homography.

[*]Equal contribution
[†]Corresponding authors

allax and foreground motion in the real world. As shown in Fig. 1(a), the two images are fully aligned using the GT homography, but the car in the red box should be a moving object that cannot be aligned by a homography.

In this work, we propose an iterative framework, which is designed to generate image pairs that satisfy both label and realism criteria [10] for supervised homography learning and learns a state-of-the-art homography estimation network with the generated dataset. Specifically, given an unlabeled source and target image (I_s and I_t) captured in real-world scenes, we use the homography and dominant plane masks of the two images estimated by the homography estimation network and a pre-trained dominant plane detection network, respectively, along with a sampled GT homography to generate a new target image $I_t^{'}$. Then, we use the I_s and $I_t^{'}$, together with the sampled GT homography, to compose a training sample of our supervised dataset. The dominant planes of I_s and $I_t^{'}$ can be fully aligned by the GT homography while the rest remain in realistic motion, making the two images satisfy both label criteria and realism criteria. As shown in Fig. 1(b), the dominant plane of the generated image pair can be fully aligned by the GT homography, while the scene parallax and dynamic objects are maintained in the foreground.

In addition, few artifacts could exist in the generated target image in early iterations, we, therefore, propose a content consistency module and a quality assessment module to eliminate the unexpected content and select high-quality image pairs for training, respectively. The selected image pairs are used to update the homography estimation network which is utilized to estimate homographies for image generation in the next iteration. During the iterative learning steps, the capability of the network is gradually improved, as well as the quality of the synthesized dataset. With our framework, any unlabeled image pairs can be used to generate training samples, thus addressing the lack of qualified datasets in supervised homography learning, which improves the performance of supervised methods and enables them to be better generalized to real-world scenes. In summary, our main contributions are threefold:

- We propose an iterative deep framework to generate a realistic dataset from unlabeled real-world image pairs for supervised homography learning and simultaneously obtain a high-precision network based on the generated dataset.
- We propose a content consistency module and a quality assessment module, achieving the elimination of unexpected content and the selection of qualified data for training.
- Experimental results show that our method brings noticeable image realism improvement compared to the prior dataset generation strategy and achieves state-of-the-art performance on public benchmarks.

2. Related Work

2.1. Traditional Method

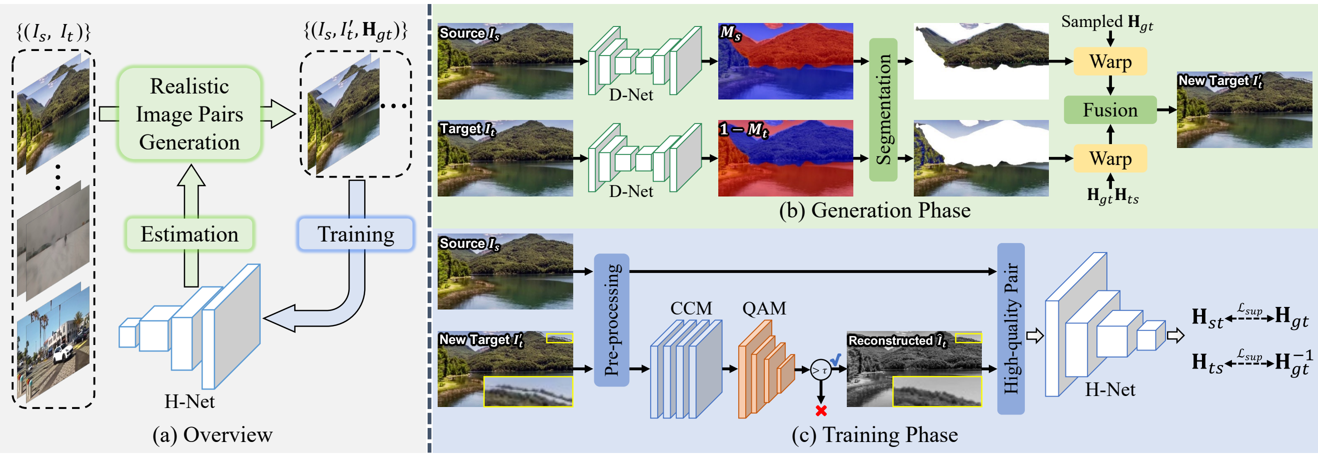
Traditional homography estimation methods can be divided into feature-based methods and optimization-based methods. Feature-based methods usually combine feature extraction and matching algorithms [26, 3, 37, 7, 34, 36] with outliner suppression approaches [9, 2], followed by solving DLT [11] to obtain homography. However, such methods would crash in challenging scenes where sufficient feature matches cannot be obtained. Optimization-based methods [5, 8] use the Lucas-Kanade algorithm [1] or calculate the sum of squared differences between two images to optimize an initialized homography iteratively. But optimization-based approaches are time-consuming and suffer from cumulative errors.

2.2. Learning-based Method

Following the development of deep image alignment methods, such as optical flow [16, 21, 10, 27, 28] and dense correspondence [39, 38], the first deep homography estimation network was proposed by [6]. Deep learning-based methods can be divided into two categories: supervised and unsupervised. The former ones [15, 35, 4] utilize synthetic image pairs that are derived from single images for training, which suffer from lacking dynamic objects and realistic scene parallax, hindering their generalization ability to real-world scenes. Unsupervised methods [30, 14, 42, 41, 12] are more robust than supervised ones thanks to their label-free training strategies. However, existing unsupervised methods are mostly optimized by minimizing the photometric distance from the warped source to the target, being sensitive to dynamic objects and homogeneous regions [38]. Besides, unsupervised methods are unstable and difficult to converge during the training phase compared to supervised ones. Therefore, our work aims to generate a realistic dataset with homography labels to optimize the network in a supervised manner and enable the network to better generalize to unseen scenes.

2.3. Deep Homography Estimation Dataset

The MS-COCO [18] and CA-unsup [42] datasets are commonly used for supervised and unsupervised homography learning. Recently, the GHOF dataset [17] is introduced for unsupervised gyroscope-guided optical flow and homography estimation. The MS-COCO dataset only contains source images, the target images are obtained by warping source images using pre-defined GT homographies. As such, the synthetic image pairs satisfy the label criteria but lack realism criteria. The CA-unsup and GHOF



$\{(I_s, I_t)\}$
Realistic Image Pairs Generation
$\{(I_s, I_t', \mathbf{H}_{gt})\}$
Estimation
Training
H-Net
(a) Overview
Source I_s
D-Net
M_s
Target I_t
D-Net
$1 - M_t$
Segmentation
Sampled $\mathbf{H}_{gt}$
Warp
Fusion
Warp
$\mathbf{H}_{gt}\mathbf{H}_{ts}$
New Target I_t'
(b) Generation Phase
Source I_s
New Target I_t'
Pre-processing
CCM
QAM
Reconstructed $\hat{I}_t$
High-quality Pair
H-Net
$\mathbf{H}_{st} \xleftrightarrow{\mathcal{L}_{sup}} \mathbf{H}_{gt}$
$\mathbf{H}_{ts} \xleftrightarrow{\mathcal{L}_{sup}} \mathbf{H}_{gt}^{-1}$
(c) Training Phase

Figure 2. The overall pipeline of our proposed framework. In the generation phase, we synthesize a new target image and form a training sample with the source image according to the dominant plane masks and homography of the original image pair estimated by D-Net and H-Net. In the training phase, we propose a content consistency module and a quality assessment module to prepare qualified image pairs for training H-Net that is adopted for image generation in the next iteration. Moreover, we perform the same pre-processing on the training pairs as in the previous method [6, 42, 41, 12], including central cropping, grayscale transformation, and normalization.

datasets collect image pairs from consecutive real-world video frames that satisfy the realism criteria but lack GT labels. In contrast, our method can synthesize realistic image pairs with labels from any unlabeled image pairs, instead of single images, to satisfy both label and realism criteria.

3. Method

3.1. Overview

The overall pipeline of our method is illustrated in Fig. 2. The core innovation of our method is that generating training data and estimating homography are mutually reinforcing. We, therefore, integrate dataset generation and network training into an iterative process, as shown in Fig. 2(a). Our method consists of two phases: the generation phase (G-phase) and the training phase (T-phase).

- **G-phase**: Given an unlabeled source image I_s and target image I_t, we generate a new target image $I_t^{'}$ according to a pre-defined homography $\mathbf{H}_{gt}$ as well as the homography and dominant plane masks of the two images estimated by the homography estimation network (H-Net) denoted as Θ_H and the dominant plane detection network (D-Net) as Θ_D, forming a training sample as $x = (I_s, I_t^{'}, \mathbf{H}_{gt})$, i.e.,

$$x = \mathrm{G}(I_s, I_t, \Theta_D, \Theta_H). \tag{1}$$

- **T-phase**: Using the generated training samples $X = \{x\}$ to update the H-Net with the help of a content consistency module (CCM) denoted as Θ_c and a quality assessment module (QAM) as Θ_q, i.e,

$$\Theta_H^{'} = \arg\min_{\theta} \mathcal{L}\left(X, \Theta_H, \Theta_c, \Theta_q\right), \tag{2}$$

where $\Theta_H^{'}$ denotes the H-Net retrained on the training samples and $\mathcal{L}$ is the learning objective.

Performing G-phase and T-phase interactively can yield a qualified dataset and a high-precision network.

3.2. Generation Phase

As shown in Fig. 2(b), we aim to generate a new target image $I_t^{'}$ according to the source image I_s and target image I_t along with a sampled homography $\mathbf{H}_{gt}$ to serves as the label between I_s and $I_t^{'}$, which is modeled by randomly selecting scaling, shearing, rotation, translation, and perspective factors from pre-defined small-baseline ranges. As a result, a training sample that satisfies both label and realism criteria [10] is generated as $(I_s, I_t^{'}, \mathbf{H}_{gt})$.

Specifically, we first adopt a dominant plane detection network (D-Net), which is formed as a UNet-like structure and pre-trained in an unsupervised manner, to estimate the dominant plane masks M_s and M_t of the I_s and I_t, and segment the dominant plane region $P_d^s = I_s \cdot M_s$ and non-dominant plane region $P_n^t = I_t \cdot (1 - M_t)$. By warping the P_d^s with $\mathbf{H}_{gt}$ to form the counterpart region of $I_t^{'}$, we ensure that their dominant planes are fully aligned. To model the realistic motion between two images, we construct the non-dominant plane region of $I_t^{'}$ from I_t. One simple approach is to directly fuse the warped P_d^s and P_n^t as

$$I_t^{'} = \mathcal{W}(P_d^s + P_n^t, \mathbf{H}_{gt}), \tag{3}$$

where $\mathcal{W}(\cdot)$ represents the warp operation. However, there exists relative motion between the two images, and the masks may not be accurate and dense enough to segment continuous and complete regions, artlessly using Eq.(3) to form the new target image would produce discontinuities in

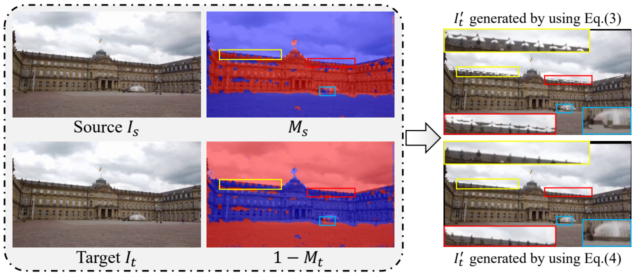

I_t' generated by using Eq.(3)
Source I_s
M_s
Target I_t
$1 - M_t$
I_t' generated by using Eq.(4)

Figure 3. Illustration of generated images by using different fusion functions. The image fused by using Eq.(3) leads to unsmooth edges and artifacts while using Eq.(4) renders a natural image.

the fusion boundary and artifacts. To solve this problem, we construct the non-dominant plane region of $I_t^{'}$ by warping the counterpart of I_t with the cumulative multiplication result of the homography from target to source $\mathbf{H}_{ts}$ and $\mathbf{H}_{gt}$, and the final $I_t^{'}$ can be obtained as

$$I_t^{'} = \mathcal{W}(P_d^s, \mathbf{H}_{gt}) + \mathcal{W}(P_n^t, \mathbf{H}_{gt}\mathbf{H}_{ts}). \tag{4}$$

As shown in Fig. 3, despite the inaccuracy of M_s and M_t, artifacts are avoided and image smoothness is guaranteed using Eq.(4).

3.3. Training Phase

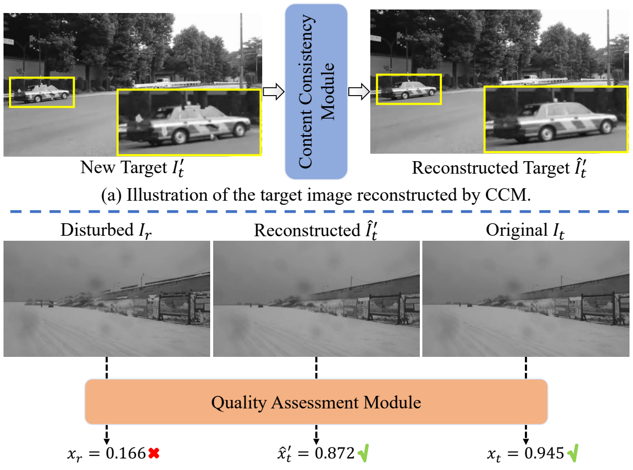
Through the G-phase, we obtain a large-scale realistic dataset for improving the homography estimation network (H-Net). As mentioned above, the quality of generated images depends on the accuracy of the estimated dominant plane masks and homography, few artifacts could exist in generated images in early iterations. To this end, we propose a content consistency module and a quality assessment module to refine the generated image and further select qualified image pairs for training, as shown in Fig. 2(c).

Content Consistency Module. The content consistency module (CCM) is designed to eliminate artifacts in $I_t^{'}$ to improve the content quality. As described in Sec. 3.2, the dominant plane of the I_s and $I_t^{'}$ can be fully aligned by the $\mathbf{H}_{gt}$ as $P_d^{t^{'}} = \mathcal{W}(P_d^s, \mathbf{H}_{gt})$, and the P_d^t and P_d^s can be aligned by the $\mathbf{H}_{ts}$ as $P_d^s = \mathcal{W}(P_d^t, \mathbf{H}_{ts})$ once the $\mathbf{H}_{ts}$ is accurate. Therefore, the dominant plane of $I_t^{'}$ can be converted into the counterpart of I_t as

$$P_d^{t^{'}} = \mathcal{W}(\mathcal{W}(P_d^t, \mathbf{H}_{ts}), \mathbf{H}_{gt}) = \mathcal{W}(P_d^t, \mathbf{H}_{gt}\mathbf{H}_{ts}). \tag{5}$$

Besides, the non-dominant plane of $I_t^{'}$ is obtained by warping the non-dominant plane of I_t with the accumulated homography $\mathbf{H}_{gt}\mathbf{H}_{ts}$. Therefore, where $\mathbf{H}_{ts}$ is extremely accurate, the $I_t^{'}$ can be obtained by warping I_t with the accumulated homography as $I_t^{'} = \mathcal{W}(I_t, \mathbf{H}_{gt}\mathbf{H}_{ts})$, which is the assumption behind the CCM we designed. To achieve this, a content consistency loss $\mathcal{L}_{ccl}$ is proposed as

$$\mathcal{L}_{ccl} = |\mathcal{W}(\mathcal{F}(\hat{I}_t^{'}), (\mathbf{H}_{gt}\mathbf{H}_{ts})^{-1}) - \mathcal{F}(I_t)|_1, \tag{6}$$


New Target I_t'
Content Consistency Module
Reconstructed Target $\hat{I}_t'$

(a) Illustration of the target image reconstructed by CCM.

Disturbed I_r
Reconstructed $\hat{I}_t'$
Original I_t
Quality Assessment Module
x_r = 0.166 ✗
$\hat{x}_t'$ = 0.872 ✓
x_t = 0.945 ✓

(b) Illustration of the quality scores estimated by QAM.

Figure 4. Illustration of the capabilities of our proposed modules. (a) shows that the artifacts are successfully removed by applying our CCM, yielding high-quality images for training. (b) shows the quality scores of different images estimated by QAM, where high-quality images are used for training and inversely rejected.

where $\hat{I}_t^{'}$ is the reconstructed result of $I_t^{'}$ by CCM, i.e., $\hat{I}_t^{'} = \Theta_c(I_t^{'})$, and $\mathcal{F}(\cdot)$ is the feature extrator of H-Net. Instead of directly minimizing the content distance between $\hat{I}_t^{'}$ and I_t, we constrain the similarity of the features to prompt CCM to reconstruct $\hat{I}_t^{'}$ to be sharp as the real-world image, since the $\mathbf{H}_{ts}$ is not accurate enough. As shown in Fig. 4(a), the artifacts in $I_t^{'}$ have been successfully removed by CCM.

Quality Assessment Module. The quality assessment module (QAM) is designed to filter bad cases that cannot be reconstructed by CCM, further improving the quality of the training data. Specifically, the QAM is designed to be a classification network whose output $x \in \mathbb{R}$ ranges from 0 to 1, x being closer to 1 means better quality and vice versa.

To achieve this, we sample another homography to interfere with $\mathbf{H}_{ts}$ since the quality of the generated image is partially relied on the accuracy of $\mathbf{H}_{ts}$, producing a disturbing image I_r with artifacts through G-phase. The score x_r obtained by feeding I_r into QAM should be close to 0. In contrast, the original target image I_t is captured from real-world scenes, so the score x_t of I_t should be close to 1. Thus, the quality assessment loss $\mathcal{L}_{qal}$ is proposed to optimize the QAM as

$$\mathcal{L}_{qal} = \text{BCE}(x_t, 1) + \text{BCE}(x_r, 0), \tag{7}$$

where $\text{BCE}(\cdot)$ denotes the binary cross-entropy loss. As a result, when the score $\hat{x}_t^{'}$ of $\hat{I}_t^{'}$ exceeds a certain threshold τ, we consider $\hat{I}_t^{'}$ to be a high-quality image that can be used for training and inversely be rejected, as shown in Fig. 4(b).

Overall, with the help of CCM and QAM, qualified training pairs can be generated, satisfying both realism criteria and supervision labels to improve the H-Net. Through iterative learning, the performance of the network and the quality of the generated dataset are mutually improved.

3.4. Network Training

Since the training samples have GT labels, we can optimize the H-Net by minimizing the difference between $\mathbf{H}_{gt}$ and the estimated $\mathbf{H}_{st'}$. However, as mentioned in [6], it is non-trivial to directly estimate a tomography, we follow previous works [15, 4, 35] to use the 4 corners offset vectors $D_{\mathbf{H}_{st'}}$ and $D_{\mathbf{H}_{gt}}$ as supervision objectives, the corresponding homography could be computed by solving DLT [11] using the offset vectors. In addition, we compute the bidirectional homography as the $\mathbf{H}_{t's}$ should also be aligned with $\mathbf{H}_{gt}^{-1}$. The finally supervised loss $\mathcal{L}_{sup}$ is expressed as

$$\mathcal{L}_{sup} = |D_{\mathbf{H}_{st'}} - D_{\mathbf{H}_{gt}}|_1 + |D_{\mathbf{H}_{t's}} - D_{\mathbf{H}_{gt}^{-1}}|_1. \qquad (8)$$

The total loss $\mathcal{L}_{total}$ is expressed by combing the supervised loss, the content consistency loss, and the quality assessment loss as

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \lambda_1\mathcal{L}_{ccl} + \lambda_2\mathcal{L}_{qal}, \qquad (9)$$

where λ_1 and λ_2 are empirically set as 0.5 and 0.1.

4. Experiment

4.1. Dataset

MS-COCO. The MS-COCO dataset [18] is one of the widely used datasets in the fields of object detection [19], segmentation [29], etc. Previously supervised methods [6, 15, 35, 4] used it to prepare their training datasets by perturbing each corner of single images to produce GT homographies and warp the images using the homographies to generate training pairs.

CA-Unsup. The CA-unsup dataset [42] contains 800k training pairs and 4.2k testing pairs that are captured from five types of real-world scenes, i.e., regular (RE), low texture (LT), low light (LL), small foreground (SF), and large foreground (LF), where the last four are challenging scenes for homography estimation. The training set contains only unlabeled image pairs, but for each pair of the test set, 6∼10 equally distributed matching points are manually marked for evaluation. Based on the CA-unsup dataset, we generated a new dataset that contains the same number of image pairs satisfying both label criteria and realism criteria as the CA-unusp dataset through our proposed method, named **CA-sup**. We show some examples of our CA-sup dataset in the supplementary materials.

GHOF. The GHOF dataset [17] is designed to evaluate homography and optical flow estimation with the gyroscope readings, it contains 10k training data and 300 testing pairs under 5 different categories, including regular (RE), foggy (FOG), low light (LL), rainy (RAIN), and snowy (SNOW) scenes. Compared to the CA-unsup dataset, it contains more parallax motion variations and extreme foreground ratio which are challenging to the homography methods. The sparse 5∼8 correspondences are used to evaluate the results.

4.2. Implementation Details

Our framework consists of D-Net, CCM, QAM, and H-Net. The D-Net is pre-trained on the CA-unsup dataset in an unsupervised manner, the detailed architecture and implementation details are described in the supplementary material. We choose BasesHomo [41], which balances performance and efficiency as the backbone of H-Net, and convert its output representation into 4 corners offset vectors. The CCM and QAM are training with the H-Net for 2 iterations, and the training parameters are consistent with the official implementation of BasesHomo for each iteration.

4.3. Comparison with Existing Methods

Comparison Methods. We compare our method with three categories of existing homography estimation methods: 1) feature-based methods including SIFT [26], ORB [33], SuperPoint [7] with SuperGlue [34] (SPSG), and LoFTR [36], 2) supervised methods including DHN [6], LocalTrans [35], and IHN [4], 3) unsupervised methods including CAHomo [42], BasesHomo [41], and HomoGAN [12]. For feature-based methods, we improve them with two different outlier rejection algorithms RANSAC [9] and MAGSAC [2], respectively. For supervised methods, we employ their models pre-trained on the MS-COCO dataset [18] and the models retrained on our CA-sup dataset denoted with $*$ for evaluation.

Quantitative Comparison. We report the quantitative results of all comparison methods on the CA-unsup test set in Table 1. The $\mathcal{I}_{3\times3}$ in the first row refers to a 3 × 3 identity matrix as a "no-warping" homography for reference, of which the errors reflect the original distance between point pairs. As shown in Table 1, our method achieves state-of-the-art performance in all categories and outperforms the best existing method HomoGAN by 12.82%, with the points matching error (PME) reduced from 0.39 to 0.34. The feature-based methods can perform well in regular (RE) scenes, as sufficient matching points can be obtained, while our methods and HomoGAN reduce the error by 0.08 compared to SIFT+RANSAC. More specifically, our method produces a PME of 0.216 which is lower than 0.222 of HomoGAN. The small foreground (SF) and large foreground (LF) scenes contain dynamic objects that affect the accuracy of estimated homography, and our method has the lowest PME in these two categories, even when compared to the methods utilizing outlier removal masks for robust estimation, i.e., CAHomo and HomoGAN. In low texture (LT) and low light (LL) scenes, most learning-based

1)	Methods	AVG	RE	LT	LL	SF	LF
2)	$\mathcal{I}_{3\times3}$	6.70 (+1617.95%)	7.75 (+3422.73%)	7.65 (+1765.85%)	7.21 (+1164.91%)	7.53 (+1611.36%)	3.39 (+993.55%)
3)	SIFT [26]+RANSAC [9]	1.41 (+261.54%)	0.30 (+36.36%)	1.34 (+226.83%)	4.03 (+607.02%)	0.81 (+84.09%)	0.57 (+83.87%)
4)	SIFT [26]+MAGSAC [2]	1.34 (+243.59%)	0.31 (+40.91%)	1.72 (+319.51%)	3.39 (+494.74%)	0.80 (+81.82%)	0.47 (51.61%)
5)	ORB [33]+RANSAC [9]	1.48 (+279.49%)	0.85 (+286.36%)	2.59 (+531.71%)	1.67 (+192.98%)	1.10 (+150.00%)	1.24 (+300.00%)
6)	ORB [33]+MAGSAC [2]	1.69 (+333.33%)	0.97 (+340.91%)	3.34 (+714.63%)	1.58 (+177.19%)	1.15 (+161.36%)	1.40 (+351.61%)
7)	SPSG [7, 34]+RANSAC [9]	0.71 (+82.05%)	0.41 (+86.36%)	0.87 (+112.20%)	0.72 (+26.32%)	0.80 (+81.81%)	0.75 (+141.94%)
8)	SPSG [7, 34]+MAGSAC [2]	0.63 (+61.54%)	0.36 (+63.64%)	0.79 (+92.68%)	0.70 (+22.81%)	0.71 (+61.36%)	0.70 (+125.81%)
9)	LoFTR [36]+RANSAC [9]	1.44 (+269.23%)	0.56 (+154.55%)	2.70 (+558.54%)	1.36 (+138.60%)	1.05 (+138.64%)	1.52 (+390.32%)
10)	LoFTR [36]+MAGSAC [2]	1.39 (+256.41%)	0.55 (+150.00%)	2.57 (+526.83%)	1.33 (+133.33%)	1.05 (+138.64%)	1.41 (+354.84%)
11)	CAHomo [42]	0.88 (+125.64%)	0.73 (+231.82%)	1.01 (+146.34%)	1.03 (+80.70%)	0.92 (+109.09%)	0.70 (+125.81%)
12)	BasesHomo [41]	0.50 (+28.21%)	0.29 (+31.82%)	0.54 (+31.71%)	0.65 (+14.04%)	0.61 (+38.64%)	0.41 (+32.26%)
13)	HomoGAN [12]	0.39 (+0.00%)	0.22 (+0.00%)	0.41 (+0.00%)	0.57 (+0.00%)	0.44 (+0.00%)	0.31 (+0.00%)
14)	DHN [6]	2.87 (+635.90%)	1.51 (+586.36%)	4.48 (+992.68%)	2.76 (+384.21%)	2.62 (+495.45%)	3.00 (+867.74%)
15)	LocalTrans [35]	4.21 (+978.26%)	4.09 (+1757.59%)	4.84 (+1081.14%)	4.55 (+697.98%)	5.30 (+1105.20%)	2.25 (+624.29%)
16)	IHN [4]	4.67 (+1097.44%)	4.85 (+2104.55%)	5.54 (+1251.22%)	5.10 (+794.74%)	5.04 (+1045.45%)	2.84 (+816.13%)
17)	DHN* [6]	1.89 (+384.95%)	1.21 (+451.82%)	2.13 (+418.54%)	2.33 (+307.91%)	1.72 (+291.23%)	2.07 (+567.81%)
18)	LocalTrans* [35]	1.76 (+352.03%)	1.25 (+467.41%)	2.15 (+423.85%)	1.90 (+232.61%)	2.25 (+410.41%)	1.28 (+311.77%)
19)	IHN* [4]	1.19 (+205.54%)	0.72 (+227.09%)	1.74 (+324.93%)	1.18 (+107.42%)	1.30 (+196.52%)	1.01 (+225.58%)
20)	Ours	**0.34 (-12.82%)**	**0.22 (+0.00%)**	**0.35 (-14.63%)**	**0.44 (-22.81%)**	**0.42 (-4.55%)**	**0.29 (-6.45%)**

Table 1. The point matching errors (PME) of our method and all comparison methods on the CA-unsup [42] test set. The best and second-best results are highlighted in **bold** and underlined. The percentages in the parentheses indicate the relative change in comparison to the second-best results. SPSG indicates SuperPoint with SuperGlue, and $*$ denotes the methods are retrained on our CA-sup dataset.

1)	Methods	AVG	RE	FOG	LL	RAIN	SNOW
2)	$\mathcal{I}_{3\times3}$	6.33	4.94	7.24	8.09	5.48	5.89
3)	SIFT [26]+MAGSAC [2]	3.90	**0.64**	4.01	9.77	0.70	4.40
4)	ORB [33]+RANSAC [9]	15.14	6.92	31.27	27.80	1.82	7.90
5)	SPSG [7, 34]+MAGSAC [2]	3.28	3.41	1.46	7.61	0.75	3.19
6)	LoFTR [36]+MAGSAC [2]	2.55	2.76	1.16	5.34	0.52	2.98
7)	CAHomo [42]	3.87	4.10	3.84	6.99	1.27	3.17
8)	BasesHomo [41]	2.28	2.02	1.43	4.90	0.78	2.29
9)	HomoGAN [12]	1.95	1.73	**0.60**	**3.95**	0.47	3.02
10)	DHN [6]	6.61	6.04	6.02	7.68	6.99	6.32
11)	LocalTrans [35]	5.72	4.06	6.49	5.95	5.78	6.34
12)	IHN [4]	8.17	7.10	8.71	9.34	6.57	9.13
13)	DHN* [6]	3.01	1.92	3.94	4.54	1.98	2.66
14)	LocalTrans* [35]	2.89	1.78	4.27	4.59	1.37	2.43
15)	IHN* [4]	2.59	2.21	3.05	4.70	0.98	2.03
16)	Ours	**1.72**	1.60	0.88	4.42	**0.43**	**1.28**

Table 2. The point matching errors (PME) of our method and comparison methods on the GHOF [17] test set. The best and second-best results are highlighted in **bold** and underlined.

methods are more robust due to their keypoint-free estimation strategy, especially unsupervised ones. However, homogeneous region [38] dominates a large portion of the image in the LT and LL scenes, where the photometric loss between the two images is minor, so the performance of unsupervised methods in these two categories is not as well as in others. In contrast, our method outperforms the unsupervised methods in the LT and LL with errors reduced by at least 14.63% and 22.81% respectively, thanks to our photometric-free fully supervised learning strategy.

Moreover, the generalization ability of supervised methods has been widely criticized. Therefore, we further evaluate feature-based methods that performed well on the CA-unsup dataset, learning-based methods, and our method on the GHOF test set to prove the effectiveness of our method. As shown in Table 2, our method achieves the best results in RAIN and SNOW scenes and the second-best results in RE, FOG and LL scenes, resulting in the lowest average PME. For regular scenes, even though feature-based methods can extract sufficient high-quality feature matches, we still outperform most of them. In addition, our CA-sup dataset does not contain images captured in the FOG and RAIN scenes, while our method is capable of generalizing these scenes, which proves the superiority of our method. On the other hand, previous supervised methods cannot generalize well to real-world scenes since their training data lack realistic motion. After retraining on our CA-sup dataset, their performance has been significantly improved, which further proves the effectiveness of our framework.

Qualitative Comparison. In Fig. 5 and Fig. 6, we provide the qualitative results of our method and competitive methods on the CA-unsup and GHOF test sets, respectively. Fig. 5(a) and (b) are from LF and SF scenes with dynamic objects, and our method produces more accurate results than feature-based and unsupervised methods which are sensitive to moving objects, as highlighted in the red and green boxes. Fig. 5(c) is from the LT scene, where feature-based methods all fail due to the lack of sufficient correspondence. On the contrary, learning-based methods perform well, but they are still not competitive as ours. In Fig. 6, we provide the comparison results in RAIN and FOG scenes of the GHOF test set, where our training set does not contain images captured in such scenes. As mentioned above, while unsupervised methods have better generalization ability in unseen scenes, we prove that supervised methods can also achieve satisfactory results after training on our generated dataset. For example, in Fig. 6(a), the pre-

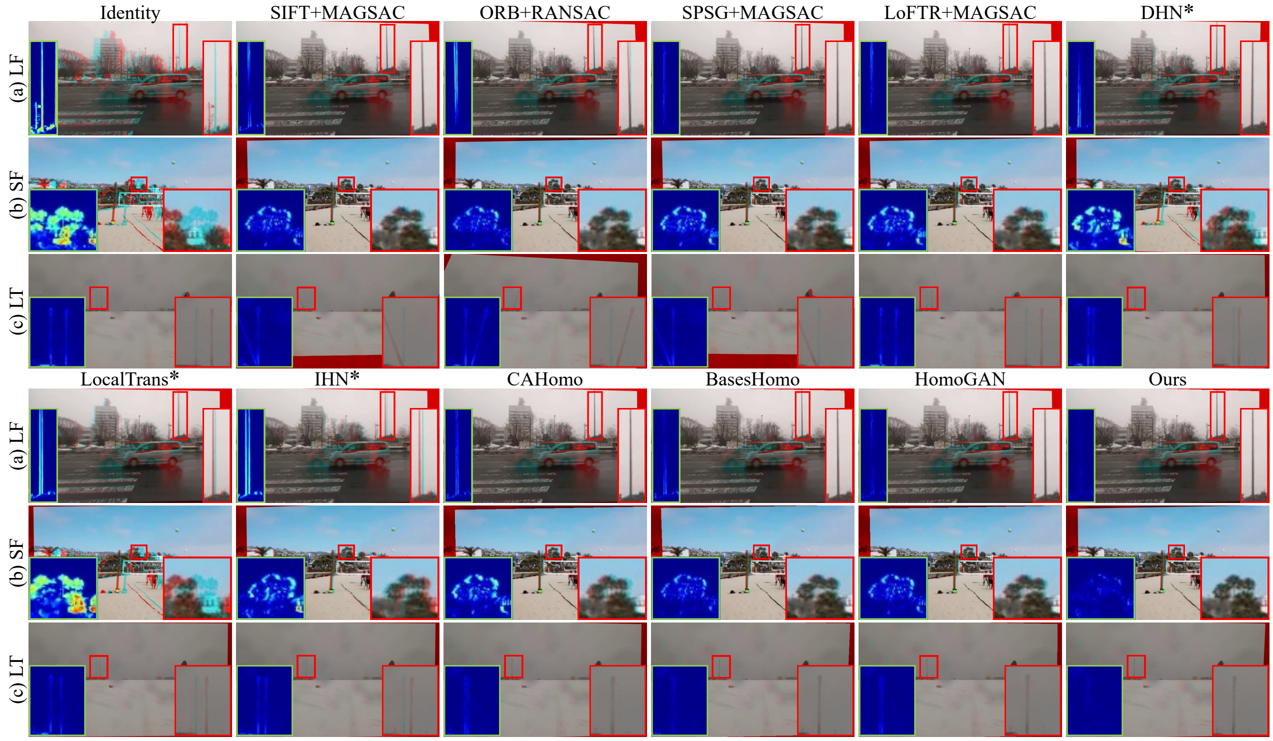


Figure 5. Qualitative results of our method and other competitive methods on the CA-unsup [42] test set. The images are generated by superimposing the warped source images on the target image. Error-prone regions are highlighted with red boxes, and the green boxes show the content difference between the two images in the error-prone regions. Best viewed by zooming in.

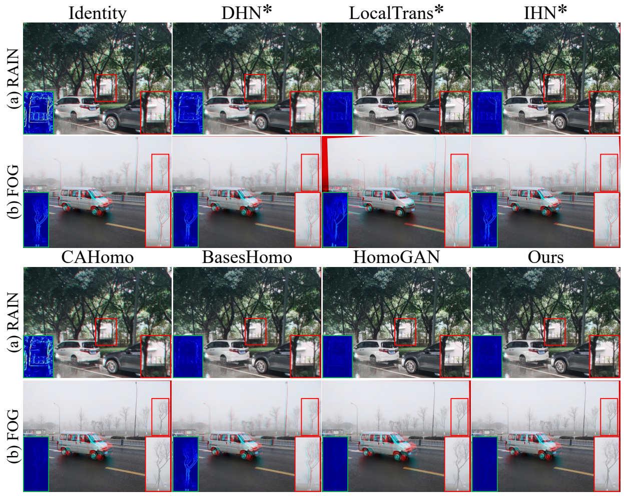


Figure 6. Qualitative results of our method and other competitive methods on the GHOF [17] test set. Best viewed by zooming in.

vious supervised methods can even obtain more accurate results than CAHomo, and our method performs the best.

Comparison with Dataset Generation Methods. We use images from the CA-unsup dataset to generate a new dataset named CA-sup† using the previous dataset generation strategy [6], and retrain H-Net on it for comparison.

	Dataset	AVG	RE	LT	LL	SF	LF
DHN [6]	CA-sup†	1.37	0.79	1.70	1.28	1.23	1.83
Ours	CA-sup	0.34	0.22	0.35	0.44	0.42	0.29

Table 3. Comparison with previous dataset generation method. We use the same source images to generate the training dataset and train the same network for comparison.

Quantitative results on the CA-unsup test set are illustrated in Table 3. The previous strategy generates image pairs from single images that meet label criteria only, causing unsatisfactory performance. In contrast, the image pairs contained in our CA-sup dataset can help the network to learn real-world motion representations, proving the importance of realistic scene motion for estimating homography.

Robustness Evaluation. Furthermore, we evaluate the robustness of all comparison methods on the CA-unsup test set by setting thresholds to calculate the proportion of inlier predictions. As shown in Fig. 7, we plot a series of curves where axis $X \in [0.1, 3.0]$ is the threshold and axis $Y \in [0.0, 1.0]$ is the percentage of inliers, higher position of curves represents better robustness. With a threshold of 0.1, our outlier percentage is 6.9% higher than the second-best method (15.8% vs. 8.9%), and our method achieves a proportion of 99.9% with a threshold of 3.0.

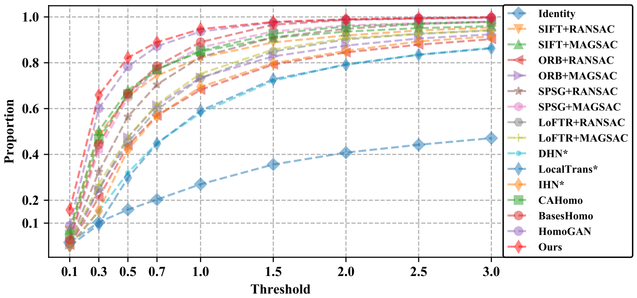


Figure 7. The proportion of inliers of our method and all comparison methods under various thresholds. The higher position of curves represents better robustness.

4.4. Ablation Study

In this section, we conduct a series of ablation studies to measure the impact of different component choices of our method. We use the CA-sup dataset for training, and quantitative results on the CA-unsup test set are illustrated in Table 4, detailed experiment settings are discussed below. For more ablation experiments, please refer to the supplementary material.

Dataset Generation Strategy. We conduct an experiment to evaluate the effectiveness of our dataset generation strategy by using G-phase only to generate training samples to train the network, denoted as Baseline in row 1 of Table 4. Even without using CCM and QAM, our method achieves state-of-the-art performance compared to the existing methods. Moreover, the homography estimation network, i.e., BasesHomo, trained on our generated dataset achieves a 24% performance gain compared to the network trained with the unsupervised learning strategy, further proving the superiority of our dataset generation strategy.

Module Effectiveness. We conduct experiments to evaluate the effectiveness of our proposed CCM and QAM by removing them separately from the overall framework. As shown in rows 2-3 of Table 4, CCM and QAM can help to generate and select high-quality image pairs for training, achieving error reduction of 0.03 and 0.01, respectively, compared to the Baseline in row 1 using none of them. By using both CCM and QAM in the training phase, our method can achieve the best performance.

Different Homography Estimation Networks. As described in Sec. 4.2, we choose BasesHomo as the backbone of our H-Net. In rows 4-5 of Table 4, we further switch the backbone to CAHomo and HomoGAN, denoted as Real-CA and Real-GAN, respectively, to prove that our method also works not only on a specific architecture. Compared with the unsupervised results in rows 11 and 13 of Table 1, CAHomo obtains performance gain by 50% with our framework and HomoGAN reduces the error by 0.09, proving the generalization of our proposed framework.

Iteration Times. The quality of generated dataset and network performance are mutually reinforcing through iterative learning. As shown in rows 6-8 of Table 4, the performance of H-Net improves with iteration increases, while higher-precision H-Net generates more natural images, as shown in Fig. 8, and high-quality images can improve performance. When the homography estimated by H-Net is accurate enough, there would be no artifacts and boundary discontinuities in the generated images. Notably, "0 iteration" indicates that the image is generated by using the homography estimated by unsupervised BasesHomo. However, a certain upper limit exists in our method and it will converge after several iterations. The performance gain of 3 iterations compared to 2 iterations, i.e., default, is not significant, and more iterations are not cost-effective compared to the resources spent per iteration.

		AVG	RE	LT	LL	SF	LF
1)	Baseline	0.38	0.23	0.40	0.50	0.45	0.33
2)	w/o CCM	0.37	0.23	0.36	0.48	0.44	0.34
3)	w/o QAM	0.35	0.23	0.36	0.45	0.42	0.30
4)	Real-CA [42]	0.44	0.27	0.45	0.52	0.53	0.43
5)	Real-GAN [12]	0.30	0.18	0.30	0.39	0.37	0.28
6)	1 iteration	0.38	0.24	0.40	0.46	0.48	0.33
7)	3 iterations	0.33	0.21	0.34	0.44	0.40	0.28
8)	Default	0.34	0.22	0.35	0.44	0.42	0.29

Table 4. Results of ablation studies, please refer to the text for more details. w/o denotes without.

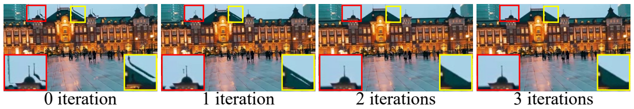

0 iteration 1 iteration 2 iterations 3 iterations

Figure 8. Illustration of the generated image in different iterations.

5. Conclusion

We have presented an iterative deep framework that aims to generate a realistic dataset for supervised homography learning and obtain a high-precision network. We use the dominant plane masks estimated by D-Net and the homography estimated by H-Net to generate training pairs, and the CCM and QAM are proposed to prepare high-quality training data to update the H-Net. The trained network and the generated data can be improved iteratively, yielding a qualified dataset as well as a state-of-the-art network. Extensive experiments demonstrate the superiority of our method and the effectiveness of our newly proposed components, the performance of existing methods can be also improved using our dataset for training.

Acknowledgements. This work is supported by Sichuan Science and Technology Program of China under grant No.2023NSFSC0462.

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. 2

[2] Daniel Barath, Jiri Matas, and Jana Noskova. Magsac: marginalizing sample consensus. In *Proc. CVPR*, pages 10197–10205, 2019. 1, 2, 5, 6

[3] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Proc. ECCV*, pages 404–417, 2006. 1, 2

[4] Si-Yuan Cao, Jianxin Hu, Zehua Sheng, and Hui-Liang Shen. Iterative deep homography estimation. In *Proc. CVPR*, pages 1879–1888, 2022. 2, 5, 6

[5] Che-Han Chang, Chun-Nan Chou, and Edward Y Chang. Clkn: Cascaded lucas-kanade networks for image alignment. In *Proc. CVPR*, pages 2213–2221, 2017. 2

[6] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Deep image homography estimation. *arXiv preprint arXiv:1606.03798*, 2016. 1, 2, 3, 5, 6, 7

[7] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proc. CVPRW*, pages 224–236, 2018. 2, 5, 6

[8] Tianjiao Ding, Yunchen Yang, Zhihui Zhu, Daniel P Robinson, René Vidal, Laurent Kneip, and Manolis C Tsakiris. Robust homography estimation via dual principal component pursuit. In *Proc. CVPR*, pages 6080–6089, 2020. 2

[9] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 1, 2, 5, 6

[10] Yunhui Han, Kunming Luo, Ao Luo, Jiangyu Liu, Haoqiang Fan, Guiming Luo, and Shuaicheng Liu. Realflow: Em-based realistic optical flow dataset generation from videos. In *Proc. ECCV*, pages 288–305, 2022. 2, 3

[11] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 1, 2, 5

[12] Mingbo Hong, Yuhang Lu, Nianjin Ye, Chunyu Lin, Qijun Zhao, and Shuaicheng Liu. Unsupervised homography estimation with coplanarity-aware gan. In *Proc. CVPR*, pages 17663–17672, 2022. 2, 3, 5, 6, 8

[13] Hai Jiang, Haipeng Li, Yuhang Lu, Songchen Han, and Shuaicheng Liu. Semi-supervised deep large-baseline homography estimation with progressive equivalence constraint. In *Proc. AAAI*, pages 1024–1032, 2023. 1

[14] Dewi Endah Kharismawati, Hadi Ali Akbarpour, Rumana Aktar, Filiz Bunyak, Kannappan Palaniappan, and Toni Kazic. Cornet: Unsupervised deep homography estimation for agricultural aerial imagery. In *Proc. ECCV*, pages 400–417, 2020. 2

[15] Hoang Le, Feng Liu, Shu Zhang, and Aseem Agarwala. Deep homography estimation for dynamic scenes. In *Proc. CVPR*, pages 7652–7661, 2020. 2, 5

[16] Haipeng Li, Kunming Luo, and Shuaicheng Liu. Gyroflow: Gyroscope-guided unsupervised optical flow learning. In *Proc. ICCV*, pages 12869–12878, 2021. 2

[17] Haipeng Li, Kunming Luo, Bing Zeng, and Shuaicheng Liu. Gyroflow+: Gyroscope-guided unsupervised deep homography and optical flow learning. *arXiv preprint arXiv:2301.10018*, 2023. 2, 5, 6, 7

[18] 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 *Proc. ECCV*, pages 740–755, 2014. 2, 5

[19] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *International Journal of Computer Vision*, 128(2):261–318, 2020. 5

[20] Shuaicheng Liu, Yuhang Lu, Hai Jiang, Nianjin Ye, Chuan Wang, and Bing Zeng. Unsupervised global and local homography estimation with motion basis learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(06):7885–7899, 2023. 1

[21] Shuaicheng Liu, Kunming Luo, Ao Luo, Chuan Wang, Fanman Meng, and Bing Zeng. Asflow: Unsupervised optical flow learning with adaptive pyramid sampling. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4282–4295, 2022. 2

[22] Shuaicheng Liu, Nianjin Ye, Chuan Wang, Kunming Luo, Jue Wang, and Jian Sun. Content-aware unsupervised deep homography estimation and its extensions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(03):2849–2863, 2023. 1

[23] Shuaicheng Liu, Lu Yuan, Ping Tan, and Jian Sun. Bundled camera paths for video stabilization. *ACM Transactions on Graphics*, 32(4):1–10, 2013. 1

[24] Zhen Liu, Wenjie Lin, Xinpeng Li, Qing Rao, Ting Jiang, Mingyan Han, Haoqiang Fan, Jian Sun, and Shuaicheng Liu. Adnet: Attention-guided deformable convolutional network for high dynamic range imaging. In *Proc. CVPRW*, pages 463–470, 2021. 1

[25] Zhen Liu, Yinglong Wang, Bing Zeng, and Shuaicheng Liu. Ghost-free high dynamic range imaging with context-aware transformer. In *Proc. ECCV*, pages 344–360, 2022. 1

[26] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 1, 2, 5, 6

[27] Ao Luo, Fan Yang, Xin Li, and Shuaicheng Liu. Learning optical flow with kernel patch attention. In *Proc. CVPR*, pages 8906–8915, 2022. 2

[28] Ao Luo, Fan Yang, Kunming Luo, Xin Li, Haoqiang Fan, and Shuaicheng Liu. Learning optical flow with adaptive graph reasoning. In *Proc. AAAI*, pages 1890–1898, 2022. 2

[29] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3523–3542, 2022. 5

[30] Ty Nguyen, Steven W Chen, Shreyas S Shivakumar, Camillo Jose Taylor, and Vijay Kumar. Unsupervised deep homography: A fast and robust homography estimation model. *IEEE Robotics and Automation Letters*, 3(3):2346–2353, 2018. 2

[31] Lang Nie, Chunyu Lin, Kang Liao, Meiqin Liu, and Yao Zhao. A view-free image stitching network based on global homography. *Journal of Visual Communication and Image Representation*, 73:102950, 2020. 1

[32] Lang Nie, Chunyu Lin, Kang Liao, and Yao Zhao. Learning edge-preserved image stitching from multi-scale deep homography. *Neurocomputing*, 491:533–543, 2022. 1

[33] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *Proc. ICCV*, pages 2564–2571, 2011. 5, 6

[34] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proc. CVPR*, pages 4938–4947, 2020. 2, 5, 6

[35] 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 *Proc. ICCV*, pages 14890–14899, 2021. 2, 5, 6

[36] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proc. CVPR*, pages 8922–8931, 2021. 2, 5, 6

[37] Yurun Tian, Xin Yu, Bin Fan, Fuchao Wu, Huub Heijnen, and Vassileios Balntas. Sosnet: Second order similarity regularization for local descriptor learning. In *Proc. CVPR*, pages 11016–11025, 2019. 2

[38] Prune Truong, Martin Danelljan, and Radu Timofte. Glu-net: Global-local universal network for dense flow and correspondences. In *Proc. CVPR*, pages 6258–6268, 2020. 2, 6

[39] Prune Truong, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning accurate dense correspondences and when to trust them. In *Proc. CVPR*, pages 5714–5724, 2021. 2

[40] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Deep high dynamic range imaging with large foreground motions. In *Proc. ECCV*, pages 117–132, 2018. 1

[41] Nianjin Ye, Chuan Wang, Haoqiang Fan, and Shuaicheng Liu. Motion basis learning for unsupervised deep homography estimation with subspace projection. In *Proc. ICCV*, pages 13117–13125, 2021. 2, 3, 5, 6

[42] Jirong Zhang, Chuan Wang, Shuaicheng Liu, Lanpeng Jia, Nianjin Ye, Jue Wang, Ji Zhou, and Jian Sun. Content-aware unsupervised deep homography estimation. In *Proc. ECCV*, pages 653–669, 2020. 2, 3, 5, 6, 7, 8

[43] Zhuofan Zhang, Zhen Liu, Bing Zeng, and Shuaicheng Liu. Minimum latency deep online video stabilization. *arXiv preprint arXiv:2212.02073*, 2022. 1