Abstract

Robust homography estimation between two images is a fundamental task which has been widely applied to various vision applications. Traditional feature based methods often detect image features and fit a homography according to matched features with RANSAC outlier removal. However, the quality of homography heavily relies on the quality of image features, which are prone to errors with respect to low light and low texture images. On the other hand, previous deep homography approaches either synthesize images for supervised learning or adopt aerial images for unsupervised learning, both ignoring the importance of depth disparities in homography estimation. Moreover, they treat the image content equally, including regions of dynamic objects and near-range foregrounds, which further decreases the quality of estimation. In this work, to overcome such problems, we propose an unsupervised deep homography method with a new architecture design. We learn a mask during the estimation to reject outlier regions. In addition, we calculate loss with respect to our learned deep features instead of directly comparing the image contents as did previously. Moreover, a comprehensive dataset is presented, covering both regular and challenging cases, such as poor textures and non-planar interferences. The effectiveness of our method is validated through comparisons with both feature-based and previous deep-based methods. Code will be soon available at Github.

1. Introduction

A homography is the fundamental image alignment model that is wildly used in computer vision [13]. It can align images taken from different perspectives, as long as the images undergo a pure rotational motion or the scene is close to a planar surface. It has been wildly used in vision applications such as image mosaicing [5], monocular SLAM [27], Co-SLAM [45], image stitching [4], video stitching [12], augmented reality [33], and camera calibration [44].

Homography alignment holds a pivot position among other image registration techniques, such as content preserving warp (CPW) that adopts mesh warps to compensate non-planer motions [21] and optical flow for dynamic objects or depth discontinuities [17]. Even in such difficult cases, it is quite helpful to align images first by a homography before more advanced models. Moreover, for scenes in which objects are far from the viewers, they can be considered as planer surfaces, and thus is applicable for a single homography model.

Homography estimation in traditional approaches can be divided into two categories, one follows the Lucas-Kanade algorithm [2] and the other requires matched image feature points [24]. In the feature-based methods, a set of fea-
ture correspondences are obtained, the homography is estimated using DLT method [13] with RANSAC outlier rejection [10]. Compared with direct methods, feature-based methods have achieved better performance. However, the quality of feature-based methods highly rely on the quality of image features. Estimation could be inaccurate when the number of matched points were insufficient or when the distribution of features were poor. High quality features should evenly distributed and cover the entire image. However, this is challenging due to the existence of textureless regions (e.g., blue sky and white wall), repetitive patterns and illumination variations.

Figure 1 shows some examples that compare our deep homography with traditional SIFT and RANSAC method. For images with rich textures (Figure 1 (a)), our method performs equally well as feature-based method. However, for images suffering from textureless regions (Figure 1 (c)), limited number of feature points can be extracted, leading to problematic homography fitting. In contrast, our deep solution is more robust under this situation. Moreover, it is challenging to conduct correct RANSAC for scenes containing large moving foreground (Figure 1 (b)) or two dominate planes (Figure 1 (d)). The failure of RANSAC leads to the failure of homography. Our deep solution is free from such problems.

DeTone et al. proposed the first deep homography approach [7]. The method takes two images as input and produces a homography from the source image to the target image. It requires ground truth homography to supervise the training. Therefore, a random homography is applied to the source image to generate the target image, forming the training pair. However, the training data generated by homography warping cannot reflect the real depth disparities. As such, the performance of DeTone et al. on real images is unsatisfactory. To solve this problem, Nguyen et al. [28] proposed an unsupervised approach which minimized the photometric loss on real image pairs. However, there are two problems. First, the loss calculated with respect to intensity is less effective than that in the feature space. Second, image regions are considered equally, ignoring the effect of ‘RANSAC’. Some image regions, such as moving objects or non-planer objects, should be excluded during the loss calculation, without which the outlier regions would decrease the estimation accuracy. Therefore, Nguyen et al. has to work on aerial images that is far away from the camera to minimize the influence of depth variations of parallax.

In this work, we propose an unsupervised approach with a new architecture for content awareness learning. In particular, we learn a content mask to reject outlier regions to mimic the traditional RANSAC procedure. To realize this, we introduce a novel triple loss for the effective optimization. Moreover, instead of comparing intensity values directly, we calculate loss with respect to our learned deep features, which is more effective. In addition, we introduce a comprehensive homography dataset, within which the testing set contains manually labeled ground-truth point matches for the purpose of quantitative comparison. The dataset consisted of 5 categories, including regular, low-texture, low-light, small-foreground, and large-foreground scenes. We show the advantages of our method over both traditional feature-based approaches and previous deep-based solutions. In summary, our main contributions are:

- A new unsupervised network structure that enables content-aware robust homography estimation from two images.
- A triple loss designed for training the network, so that a deep feature map for alignment and a mask highlighting the alignment inliers could be learned.
- A comprehensive dataset covers various scenes for both training and testing.

2. Related Work

Traditional homography. A homography is a $3 \times 3$ matrix which compensates plane motions between two images. It consists of 8 degree of freedom, with 2 for scale, 2 for translation, 2 for rotation and 2 for perspective [13]. To solve a homography, traditional approaches often detect and match image features, e.g., SIFT [24], SURF [3] and ORB [32]. Two sets of correspondences were established between two images, following which robust estimation is adopted, such as RANSAC [24] and IRLS [16], for the outlier rejection during the model estimation.

A homography can also be solved directly without image features. The direct methods, such as seminal Lucas-Kanade algorithm [25], calculates sum of squared differences (SSD) between pixels from two images. The differences guide the shift and warp of the images, yielding homography updates. A random initialized homography is optimized in this way iteratively [2]. Moreover, the SSD can be replaced with enhanced correlation coefficient (ECC) for the robustness [9].

Deep homography. Following the success of various deep image alignment methods, such as optical flow [39, 17], dense matching [31] and deep features [1], the deep homography was first proposed by [7] in 2016. The network takes source and target images as input and outputs 4 displacement vectors at 4 image corners of source image, which then yields the homography. It used ground-truth homography to supervise the training. However, the training images with GT homography is generated without depth disparity. To overcome such issue, [28] proposed an unsupervised approach that computed photometric loss between two images and adopted Spatial Transform Network (STN) [18] for image warping.
Image stitching. Our approach is also related to image stitching methods. These methods are traditional methods that target at registration images under large disparities [42] for the purpose of constructing panorama [4]. The stitched images were captured under dramatic viewpoint differences. Various methods have been proposed along this direction, such as Dual-Homography for the scenes contain two dominant planes [11], As-Projective-As-Possible (APAP) [41] and MeshFlow [23] for non-rigid mesh motion compensation, Direct Photometric Alignment (DPA) for low-textured images [20], and Shape-preserving-half-projective (SPHP) for the shape rigidity at non-overlapping regions [6]. In this work, we do not target on examples of image stitching. We focus on images with sufficient overlaps, moderate viewpoint differences and reasonable depth disparities, which can be aligned under the capability of a single homography.

3. Algorithm

3.1. Network Structure

Our method is built upon convolutional neural networks. It takes two grayscale images $I_a$ and $I_b$ as input, and produces a homography matrix $H_{ab}$ from $I_a$ to $I_b$ as output. The entire structure could be divided into three modules: a feature extractor $f(\cdot)$, a mask predictor $m(\cdot)$ and a homography estimator $h(\cdot)$. $f(\cdot)$ and $m(\cdot)$ are fully convolutional networks which accepts input of arbitrary sizes, and the $h(\cdot)$ utilizes a backbone of ResNet-34 [14] and produces 8 values. Figure 2(a) illustrates the network structure.

Feature extractor. Unlike previous DNN based methods that directly utilizes the pixel values as the feature, here our network automatically learns a feature from the input for robust feature alignment. To this end, we build a FCN that takes an input of size $H \times W \times 1$, and produces a feature map of size $H \times W \times C$. For inputs $I_a$ and $I_b$, the feature extractor shares weights and produces feature maps $F_a$ and $F_b$, i.e.

$$F_a = f(I_a), \quad F_b = f(I_b) \quad (1)$$

Mask predictor. In non-planar scenes, especially those including moving objects, there exists no single homography that can align the two views. In traditional algorithm, RANSAC is widely applied to find the inliers for homography estimation, so as to solve the most approximate matrix for the scene alignment. Following the similar idea, we build a sub-network to automatically learn the inliers’ positions. Specifically, a sub-network $m(\cdot)$ learns to produce an inlier probability map or mask, highlighting the content in the feature maps that contribute much for the homography estimation. The size of the mask is the same as the size of the feature. With the masks, we further weight the features extracted by $f$ before feeding them to the homography estimator, obtaining two weighted feature maps $G_a$ and $G_b$ as,
indeed describe the fact that I fails to reflect the fact that the original images maps, i.e. solutions, where the feature extractor only produces all zero operation.

We utilize spatial transform network [18] to achieve the warping pixel location in the masks and feature maps. Here we utilize the homography matrix \( H_{ab} \) with 8 freedoms by solving a linear system. We use \( h(\cdot) \) to represent the whole process, i.e.

\[
H_{ab} = h([G_a, G_b])
\]

The backbone of the homography estimator network follows a ResNet-34 structure. It contains 34 layers of strided convolutions followed by an adaptive pooling layer, which generates fixed size (8 in our case) of feature vectors regardless of the input feature dimensions.

We list the layer details of the three modules above in Table 1. Note that, we use the sigmoid at the last layer of the mask predictor to ensure the output value ranges from [0, 1].

3.2. Triple Loss for Robust Homography Estimation

With the homography matrix \( H_{ab} \) estimated, we warp image \( I_a \) to \( I_a' \) and then further extracts its feature map as \( F_a' \). Intuitively, if the homography matrix \( H_{ab} \) is accurate enough, \( F_a' \) should be well aligned with \( F_b \), causing a low \( l_1 \) loss between them. Considering in real scenes, a single homography matrix cannot satisfy the transformation between the two views, we also normalize the \( l_1 \) loss by \( M_a' \) and \( M_b \). Here \( M_a' \) is the warped version of \( M_a \). So the loss between the warped \( I_a \) and \( I_b \) is as follows,

\[
L_n(I_a', I_b) = \frac{\sum_i M_a'M_b \cdot |F_a' - F_b|}{\sum_i M_a'M_b}
\]

where \( F_a' = f(I_a') \), \( I_a' = Warp(I_a, H_{ab}) \) and \( i \) indicates a pixel location in the masks and feature maps. Here we utilize spatial transform network [18] to achieve the warping operation.

Directly minimizing Eq. 5 may easily cause trivial solutions, where the feature extractor only produces all zero maps, i.e. \( F_a' = F_b = 0 \). In this case, the features learned indeed describe the fact that \( I_a' \) and \( I_b \) are well aligned, but it fails to reflect the fact that the original images \( I_a \) and \( I_b \) are mis-aligned. To this end, we involve another loss between \( F_a \) and \( F_b \), i.e.

\[
L(I_a, I_b) = |F_a - F_b|
\]

and further maximize it when minimizing Eq. 5. This strategy avoids the trivial solutions, and enables the network to learn a discriminative feature map for image alignment.

In practise, we swap the features of \( I_a \) and \( I_b \) and produce another homography matrix \( H_{ba} \). Following Eq. 5 we involve a loss \( L(I_a', I_b) \) between the warped \( I_b \) and \( I_a \). We also add a constraint that enforces \( H_{ab} \) and \( H_{ba} \) to be inverse. So, the optimization procedure of the network could be written as follows,

\[
\min_{m,f,h} \sum_n L_n(I_a', I_b) + \sum_n L_n(I_b', I_a) - \lambda L(I_a, I_b)
+ \mu |H_{ab}H_{ba} - I|,
\]

where \( \lambda \) and \( \mu \) are balancing hyper-parameters, and \( I \) is a 3-order identity matrix. We set \( \lambda = 2.0 \) and \( \mu = 0.01 \) in our experiments. We illustrates the loss formulations in Figure 2(b).

3.3. Unsupervised Content-Awareness Learning

As mentioned above, our network contains a sub-network \( m(\cdot) \) to predict an inlier probability map or mask. It is such designed that our network can be of content-awareness by the two-fold effects. First, we use the masks \( M_a, M_b \) to explicitly weight the features \( F_a, F_b \), so that only highlighted features could be fully fed into homography estimator \( h(\cdot) \). Meanwhile, they are also implicitly involved into the normalized \( l_1 \) distance between the warped feature \( F_a' \) and its original counterpart \( F_b \), or \( F_b' \) and \( F_a \), meaning only those regions that are really fit for alignment would be taken into account. For those areas containing low texture or moving foreground, because they are non-distinguishable or misleading for alignment, they are naturally removed for homography estimation during optimizing the triple loss as proposed. Such a content-awareness is achieved fully by an unsupervised learning scheme, without any ground-truth mask data as supervision.

To demonstrate the effectiveness of mask, We illustrate several examples in Figure 3. For example, in Figure 3(a ~ c) where the scenes contain dynamic objects, our network successfully rejects moving objects, even if the movements

| Layer No. | Type | Kernel | Stride | Channel |
|-----------|------|--------|--------|---------|
| 1         | conv | 3      | 1      | 4       |
| 2         | conv | 3      | 1      | 8       |
| 3         | conv | 3      | 1      | 1       |

| Layer No. | Type | Kernel | Stride | Channel |
|-----------|------|--------|--------|---------|
| 1         | conv | 3      | 1      | 1       |
| 2         | conv | 3      | 1      | 8       |
| 3         | conv | 3      | 1      | 16      |
| 4         | conv | 3      | 1      | 32      |
| 5         | conv | 3      | 1      | 1       |

Table 1: Network architecture of feature extractor (a), mask predictor (b) and homography estimator (c).
Figure 3: Our predicted masks for various of scenes. (a) contains complex foreground motions. (b) and (c) contains large dynamic foreground. (d) contains few textures and (e) is an night example.

Figure 4: A glace of our dataset. (a) regular examples (RE). (b) examples with low textures (LT). (c) low light examples (LL). (d) examples with foregrounds of small sizes (SF). (e) examples of large foreground (LF).

are inapparent as the fountain in (c), or the objects occupy a large space as in (a)(b). These cases are very difficult for RANSAC to find robust inliers. In particular, the most challenging case is Figure 3(a), in which the moving foregrounds are complex, including people and the train. Our method successfully locates the useful background for the homography estimation. Figure 3(d) is a low-textured example, in which the blue sky occupies half space of the image. It is challenging for traditional methods where the sky provides no features and the sea causes matching ambiguities. Our predicted mask concentrates on the horizon but with sparse weights on sea waves. Figure 3(e) is a low light example, where only visible areas contain weights as seen. We also conduct an ablation study to reveal the influence if disabling the mask prediction. As seen in Table 2, the accuracy has a significant decrease when mask is removed.

4. Experimental Results

4.1. Dataset and Implementation Details

Previously, there is no dedicated dataset that is designed to evaluate the performance of homography fitting. The supervised method [7] synthesized homographies from a single image, so it cannot reflect disparities and occlusions. The unsupervised method [28] adopted aerial images that lacks the generalization. Therefore, we propose our own dataset for comprehensive evaluations.

Our dataset contains 5 categories, including regular (RE), low-texture (LT), low-light (LL), small-foregrounds (SF), and large-foregrounds (LF) image pairs. Each category contains around 80k image pairs, thus totally 400k image pairs in the dataset. Figure 4 shows some examples. Specifically, we collect these images from 291 video clips, each of which lasts 15 ~ 20 seconds. For each frame, we randomly sample 5 frames from its 8 consecutive later frames. With respect to the testing data, we randomly choose 100 image pairs from all categories. For each pair, we manually marked 6 ~ 8 matching points for the purpose of quantitative comparison. The marked points are equally distributed on the image.

The category partition is based on the understanding of traditional homography registration. For regular examples (Figure 4(a)), image features can be extracted easily due to rich textures and the scene is flat which is friendly for a homography. With respect to low-texture and low-light examples (Figure 4(b) and (c)), only a few number of image features could be extracted, which causes troubles for traditional homography fitting. With regards to scenes containing foreground or contain dynamic objects (Figure 4(d)), the scene is no longer a plane. In such cases, a best fitting homography would align the most dominate planar structure of the scene, with other non-planar objects be excluded. This can be achieved by RANSAC outlier rejection for traditional methods, but may cause troubles for the previous two deep methods [7, 28] which treat the image content equally. The most challenging case is the scene with large foreground (Figure 4(e)), for which even the RANSAC cannot handle it easily. We will show in the subsequent experiments that our method is robust over all categories.

Our network is trained with $120k$ iterations by an Adam optimizer [19], whose parameters are set as $\lambda_c = 1.0 \times 10^{-4}$.
Supervised Unsupervised Ours Ground-truth
Supervised Unsupervised Ours Ground-truth

Figure 5: Comparison with existing approaches. Supervised [7], Unsupervised [28], Ours and Ground-truth are shown by green, yellow, blue and red rectangles, respectively. (a) A single frame synthesized example. (b) Real consecutive frames of (a). (c) An example with a dominate plane. (d) A flash and no flash example with illumination differences of two frames caused by camera flash. (e) An example with near-range foreground at corners. (f) An example with two dominate planes plus large moving foreground. (g) An example with poor textures. (h) A low light example. Please refer to webpage https://github.com/JirongZhang/DeepHomography for more examples, where we toggle the images with GIF animation for clearer illustration.

$\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 1.0 \times 10^{-8}$. The batch size is set to 64. For every $12k$ iterations, we multiply the learning rate by $0.8$. Each iteration costs approximate 1.2s and it takes nearly 40 hours to complete the entire training. The detailed network configuration with respect to feature extractor, mask predictor and homography estimator are summarized in Table 1. The implementation is based on PyTorch and the network training is performed on 4 NVIDIA RTX 2080 Ti. To augment the training data and avoid black boundaries appearing in the warped image, we randomly crop patches of size $315 \times 560$ from the original image to form $I_a$ and $I_b$.

4.2. Comparisons with Existing Methods

**Qualitative comparison.** We first compare our method with the existing two deep homography ones, i.e. the supervised estimation [7] and the unsupervised estimation [28] approaches. Figure 5 shows the results of qualitative comparison.

In Figure 5(a), we synthesized an example with a single frame. Therefore, no disparities are introduced. In such case, all methods perform equally well, as indicated by the coincidence of rectangles. However, when we test consecutive frames of the same footage (Figure 5(b)), the supervised approach fails. Because the supervised approach cannot handle large disparities as well as moving objects of the scene. Note that, except the Figure 5(a), all examples come from real different frames. In Figure 5(c), the building surface is a plane, all methods work well in this case. Interestingly, this footage contains camera flashes, where the flash causes the illumination variation across different images. We choose the flash and no flash images for alignment (Figure 5(d)). All the three methods drift from ground-truth rectangle, while our method is the closest. This indicates that our method is robust to a certain amount of illumination change. Figure 5(e) contains near-range objects at corners and Figure 5(f) contains two dominate planes with moving objects at corners. Figure 5(g) is a low texture example and Figure 5(h) is a low light example. Similarly, in both scenarios, our method is the best among other candidates.

**Quantitative comparison.** Beyond qualitative comparisons, we also verify the effectiveness of our method by comparing with two deep methods quantitatively. The comparison is based on our dataset. In particular, the testing set for each category contains ground-truth labels. For each pair of image, we manually marked 6 to 8 correspondances. We use the estimated homography to transform the source points to the target points. The averaged $l_2$ distances are recorded as an evaluation metric like in [38, 37, 8, 22, 26]. We report the performances for each category as well as the overall averaged scores in Figure 6 left. We use a $3 \times 3$ identity matrix $I_{3 \times 3}$ as a special reference homography. As seen, our method outperforms the others for all categories in the comparison.

Specifically, in Figure 6 left, our method performs best
on the regular (RE) category compared with other categories, as the difficulty for RE is the smallest. When compared with supervised method, the large improvements have been achieved with respect to low-texture (LT) and large-foreground (LF) categories, with 1.29 and 1.22 improvements, respectively, which indicates that the selection of content is crucial for these two categories. For low-texture class, only a small amount of area is informative which should be accurately identified. With regards to large-foreground class, no matter the foreground is static or dynamic, it would confuse the estimation. If the foreground is static, as it stays close to the camera, the disparity issue will be magnified. If it is dynamic, it will violate the camera motion. For both cases, the foreground must be rejected correctly to estimate accurate homography on the background. Note that, the observations above are also applied to the low-light and small foreground categories.

We further compare our method with traditional feature-based methods. The results are shown in Figure 6 right. We verified three popular features, SIFT [24], SURF [3] and ORB [32]. For regular class (RE), rich texture delivers sufficient high quality features, while both SIFT and SURF are only a little better than ours, in which SURF achieved the best performance on the RE category. For the other categories, our method significantly outperforms the others.

Note that, for LT, LL and LF categories, the traditional feature-based methods frequently fail badly. The failure is caused by various reasons, such as limited number of detected features with poor distributions, or the failure of RANSAC given raise to inlier features located both on foreground and background. Figure 7 shows some examples. This type of failure often leads to huge errors. Therefore, for examples that are totally fail, we resort back to the least $T_{3 \times 3}$ scores (Figure 6 left, $T_{3 \times 3}$), to produce reasonable values for a relatively fair comparison. Specifically, the percentage of total failure is 27%, 23%, and 33% for LT, LL, and LF categories, respectively.

**Figure 6:** Comparison with existing methods (left: with previous DNN based methods, right: with feature-based methods).

**Figure 7:** The failure examples of feature-based method. Red: ground-truth. Blue: ours. Yellow: SIFT. We also show the failure image caused by the incorrect homography of the feature-based method on the right column.

### 4.3. Ablation Studies

**Content-aware mask.** Content-aware is the most important feature of our network. Therefore, we compare the performance in the case of with and without mask. Table 2 ‘w/o. Mask’ shows the results. The results with mask are definitely better than the results without mask. It is clear that the mask plays an important role in the deep homography, just like the importance of RANSAC to the feature-based methods. The mask not only rejects dynamic regions, but also selects reliable areas for deep homography estimation.

**Triple loss.** We exam the effectiveness of our triple loss by removing the term of Eq. 6 from Eq. 7. Table 2 ‘w/o. Triple loss’ shows the result. It is clear that the triple loss
not only avoids the problem of obtaining trivial solutions, but also facilitates a better optimization.

**Backbone.** Feature backbone is another important aspect that should be studied. Here, we exam several popular backbones, including VGG [34], ResNet-18, ResNet-34 [14], and ShuffleNet [43]. As seen, the ResNet-18 achieves similar performance as our best result obtained by ResNet-34. The VGG backbone is slightly worse than ResNet-18 and ResNet-34. Interestingly, the light-weight backbone ShuffleNet-v2 achieves the performance on par with other large backbones, which indicates that our proposed method can be platted into portable or embedded systems, facilitating a wide application scope.

**Training strategy.** We adopted a separate training strategy to train our network as in [36, 15, 29], i.e. at the very beginning, train the feature extractor only without the mask predictor involved by setting the mask to all ones. With stable features have been trained from the extractor, i.e. about 60k iterations in our experiments, we finetuned the network with mask predictor involved, with a learning rate of $5 \times 10^{-5}$. To validate the effectiveness of this training strategy, we also did an experiment by training the feature extractor and mask predictor simultaneously, both from scratch. Table 2 compares the performances of two training strategies. As reported, our separate training posses better performance.

|                      | RE | LT | LL | SF | LF | Avg |
|----------------------|----|----|----|----|----|-----|
| w/o. Mask            | 1.16 | 1.85 | 1.45 | 1.38 | 1.25 | 1.42 |
| w/o. Triple loss     | 1.37 | 2.78 | 2.20 | 1.53 | 1.98 | 1.97 |
| VGG                  | 1.08 | 1.90 | 1.49 | 1.36 | 1.38 | 1.44 |
| ResNet-18            | 1.10 | 1.54 | 1.54 | 1.17 | 1.24 | 1.32 |
| ShuffleNet-v2        | 1.18 | 1.55 | 1.40 | 1.29 | 1.26 | 1.34 |
| Train from scratch   | 1.10 | 1.73 | 1.45 | 1.20 | 1.34 | 1.36 |
| Ours(ResNet-34)      | 1.02 | 1.48 | 1.29 | 1.15 | 1.20 | 1.23 |

Table 2: Ablation studies on mask, triple loss, training strategy and network backbones. Data represents the $l_2$ distances between transformed points and marked ground-truth points.

5. Conclusions

We have presented a new architecture for deep homography estimation with content-aware capability. Traditional feature based methods heavily rely on the quality of image features which are vulnerable to low-texture and low-light scenes. Large foreground also causes troubles for RANSAC outlier removal. Previous DNN based methods pay less attention to the depth disparity issue. They treat the image content equally which can be influenced by non-planar structures or dynamic objects. Our network learn a mask during the estimation to reject outlier regions for robust homography estimation. In addition, we calculate loss with respect to our learned deep features instead of directly comparing the image intensities. Moreover, we have provided a comprehensive homography dataset. The dataset have been divided into 5 categories, regular, low-texture, low-light, small-foregrounds, and large-foregrounds, to evaluate performances under different aspects. The comparison with previous methods show the effectiveness of our method.

References

[1] H. Altwaijry, A. Veit, S. J. Belongie, and C. Tech. Learning to detect and match keypoints with deep architectures. In *Proc. BMVC*, 2016. 2
[2] S. Baker and I. Matthews. Lucas-kadane 20 years on: A unifying framework. *International journal of computer vision*, 56(3):221–255, 2004. 1, 2
[3] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *Proc. ECCV*, pages 404–417, 2006. 2, 7
[4] M. Brown, D. G. Lowe, et al. Recognising panoramas. In *Proc. ICCV*, volume 3, page 1218, 2003. 1, 3
[5] D. Capel. Image mosaicing. In *Image Mosaicing and super-resolution*, pages 47–79, 2004. 1
[6] C.-H. Chang, Y. Sato, and Y.-Y. Chuang. Shape-preserving half-projective warps for image stitching. In *Proc. CVPR*, pages 3254–3261, 2014. 3
[7] D. DeTone, T. Malisiewicz, and A. Rabinovich. Deep image homography estimation. *arXiv preprint arXiv:1606.03798*, 2016. 2, 5, 6
[8] Y. Ding, C. Wang, H. Huang, J. Liu, J. Wang, and L. Wang. Frame-recurrent video inpainting by robust optical flow inference. *arXiv preprint arXiv:1905.02882*, 2019. 6
[9] G. D. Evangelidis and E. Z. Psarakis. Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 30(10):1858–1865, 2008. 2
[10] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2
[11] J. Gao, S. J. Kim, and M. S. Brown. Constructing image panoramas using dual-homography warping. In *Proc. CVPR*, pages 49–56, 2011. 3
[12] H. Guo, S. Liu, T. He, S. Zhu, B. Zeng, and M. Gabbouj. Joint video stitching and stabilization from moving cameras. *IEEE Trans. on Image Processing*, 25(11):5491–5503, 2016. 1
[13] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 1, 2
[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, 2016. 3, 8
E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and P. W. Holland and R. E. Welsch. Robust regression using iteratively reweighted least-squares. Communications in Statistics-theory and Methods, 6(9):813–827, 1977. 2

E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In in Proc. CVPR, pages 2462–2470, 2017. 1, 2

M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In Advances in neural information processing systems, pages 2017–2025, 2015. 2, 4

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014. 5

K. Lin, N. Jiang, S. Liu, L.-F. Cheong, M. Do, and J. Lu. Direct photometric alignment by mesh deformation. In in Proc. CVPR, pages 2405–2413, 2017. 3

F. Liu, M. Gleicher, H. Jin, and A. Agarwala. Content-preserving warps for 3d video stabilization. In ACM Trans. Graph. (Proc. of SIGGRAPH), volume 28, page 44, 2009. 1

J. Liu, C.-H. Wu, Y. Wang, Q. Xu, Y. Zhou, H. Huang, C. Wang, S. Cai, Y. Ding, H. Fan, et al. Learning raw image denoising with bayer pattern unification and bayer preserving augmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 0–0, 2019. 6

S. Liu, P. Tan, L. Yuan, J. Sun, and B. Zeng. Meshflow: Minimum latency online video stabilization. In in Proc. ECCV, pages 800–815, 2016. 3

D. G. Lowe. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2):91–110, 2004. 1, 2, 7

B. D. Lucas, T. Kanade, et al. An iterative image registration technique with an application to stereo vision. 1981. 2

X. Meng, X. Deng, S. Zhu, S. Liu, C. Wang, C. Chen, and B. Zeng. Mganet: A robust model for quality enhancement of compressed video. arXiv preprint arXiv:1811.09150, 2018. 6

R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos. Orb-slam: a versatile and accurate monocular slam system. IEEE Trans. on robotics, 31(5):1147–1163, 2015. 1

T. Nguyen, S. W. Chen, S. S. Shivakumar, C. J. Taylor, and V. Kumar. Unsupervised deep homography: A fast and robust homography estimation model. IEEE Robotics and Automation Letters, 3(3):2346–2353, 2018. 2, 5, 6

H. Qiu, C. Wang, H. Zhu, X. Zhu, J. Gu, and X. Han. Two-phase hair image synthesis by self-enhancing generative model. arXiv preprint arXiv:1902.11203, 2019. 8

J. Ren, Y. Hu, Y.-W. Tai, C. Wang, L. Xu, W. Sun, and Q. Yan. Look, listen and learn a lag multimodal lstm for speaker identification. In Thirtieth AAAI Conference on Artificial Intelligence, 2016. 7

J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid. Deepmatching: Hierarchical deformable dense matching. International Journal of Computer Vision, 120(3):300–323, 2016. 2

E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski. Orb: An efficient alternative to sift or surf. In in Proc. ICCV, volume 11, pages 2564–2571, 2011. 2, 7

G. Simon, A. W. Fitzgibbon, and A. Zisserman. Markerless tracking using planar structures in the scene. In in Proc. International Symposium on Augmented Reality, pages 120–128, 2000. 1

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014. 8

C. Wang, Y. Guo, J. Zhu, L. Wang, and W. Wang. Video object co-segmentation via subspace clustering and quadratic pseudo-boolean optimization in an mrf framework. IEEE Transactions on Multimedia, 16(4):903–916, 2014. 7

C. Wang, H. Huang, X. Han, and J. Wang. Video inpainting by jointly learning temporal structure and spatial details. In Proceedings of the 33th AAAI Conference on Artificial Intelligence, 2019. 8

C. Wang, J. Zhu, Y. Guo, and W. Wang. Video vectorization via tetrahedral remeshing. IEEE Transactions on Image Processing, 26(4):1833–1844, 2017. 6

Y. Wang, H. Huang, C. Wang, T. He, J. Wang, and M. Hoai. Gif2video: Color dequantization and temporal interpolation of gif images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1419–1428, 2019. 6

P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. Deepflow: Large displacement optical flow with deep matching. In in Proc. CVPR, pages 1385–1392, 2013. 2

P. Yan, G. Li, Y. Xie, Z. Li, C. Wang, T. Chen, and L. Lin. Semi-supervised video salient object detection using pseudo-labels. arXiv preprint arXiv:1908.04051, 2019. 7

J. Zaragoza, T.-J. Chin, M. S. Brown, and D. Suter. As-projective-as-possible image stitching with moving dlt. In in Proc. CVPR, pages 2339–2346, 2013. 3

F. Zhang and F. Liu. Parallax-tolerant image stitching. In in Proc. CVPR, pages 3262–3269, 2014. 3

X. Zhang, X. Zhou, M. Lin, and J. Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In in Proc. CVPR, pages 6848–6856, 2018. 8

Z. Zhang. A flexible new technique for camera calibration. IEEE Trans. on Pattern Analysis and Machine Intelligence, 22, 2000. 1

D. Zou and P. Tan. Coslam: Collaborative visual slam in dynamic environments. IEEE Trans. on Pattern Analysis and Machine Intelligence, 35(2):354–366, 2012. 1