WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 4, Number 3, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v4n3.46
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Unsupervised Deep Learning Model for Homography Estimation

Lang Zhou *

College of Electronic Information, Southwest Minzu University, ChengDu, China

ABSTRACT

This paper proposes an unsupervised deep learning model for homography estimation, addressing limitations of traditional feature-based methods and supervised learning approaches. By leveraging reprojection error as the optimization objective, the model eliminates the need for labeled data while achieving precise homography estimation. The framework comprises a feature extraction module, a feature difference module, and a homography regression network. Extensive experiments on the MS-COCO and HPatches datasets demonstrate that the proposed model achieves a mean reprojection error (MRE) of 3.67 pixels and an accuracy (ACC) of 88.3%, closely approaching the performance of supervised methods like DeepHomography while significantly outperforming classical SIFT + RANSAC. The model's lightweight design ensures efficient inference, requiring only 12ms per estimation, making it suitable for real-time applications. Ablation studies validate the effectiveness of key components such as the feature difference module and regularization loss, highlighting their contributions to performance improvement. Compared to traditional methods, the proposed approach exhibits superior robustness under varying lighting conditions, viewpoint changes, and noise interference. Moreover, it removes the dependency on labeled data, reducing application costs and barriers. This unsupervised framework presents a practical and efficient solution for homography estimation and offers potential for broader applications in multi-view geometry and 3D reconstruction tasks. This is an example abstract. It describes the content and purpose of the paper succinctly. The abstract should be single-spaced and use Times New Roman 10 pt font.

KEYWORDS

Depth homography estimation; Unsupervised; Image stitching

1. INTRODUCTION

Estimating homography from image pairs is a fundamental problem in image registration/alignment [1]. Homography estimation plays an important role in a wide range of applications such as image/video stitching [2, 3], camera calibration [4], HDR imaging [5], and SLAM [6]. It is defined as the estimation of the projective transformation between two views on the same plane in three-dimensional space [7]. The homography matrix is a 3×3 matrix with 8 degrees of freedom for scale, translation, rotation, and perspective [8]. Traditionally, homography is usually estimated by detecting and matching image features and then solving the direct linear transform (DLT) by removing outliers [9, 10]. In contrast, deep homography methods take two images as network input and directly output a homography matrix [11]. Traditional homography estimation usually requires feature point matching and geometric correction, while unsupervised homography estimation uses deep methods to avoid the need for manual feature point annotation, thereby improving efficiency and automation. Deep methods can be divided into two categories, supervised [12] and unsupervised [13, 14]. The former uses synthetic samples with ground-truth labels to train the network, while the latter directly

minimizes the photometric or feature difference between two images. Since synthetic samples cannot reflect scene parallax and dynamic objects, the generalization effect of unsupervised methods is often better than that of supervised methods.

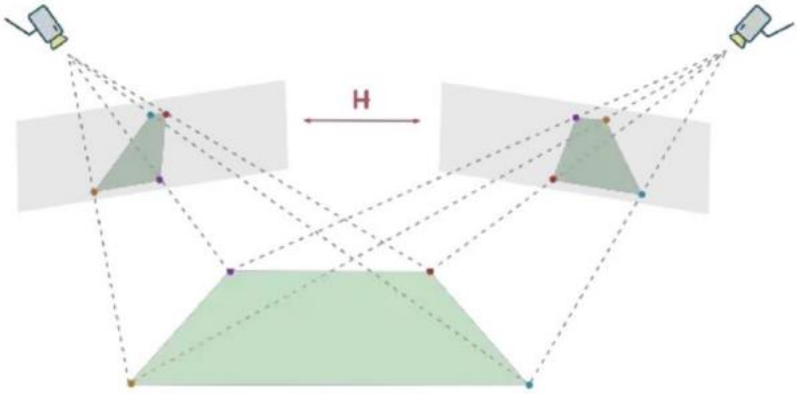

Figure 1. Illustration of the definition of homography

In unsupervised homography estimation, convolutional neural networks (CNNs) are usually used as the core model. This method abandons geometric features and instead adopts high-level semantic features, which can be adaptively learned in a data-driven mode in a supervised [15], weakly supervised [16] or unsupervised manner. The input is a pair of images. The model estimates the homography matrix by learning the transformation relationship between the images. During the training process, a loss function based on the reconstruction error is used to guide the model optimization. Specifically, the transformed image can be inversely transformed by the estimated homography matrix, and the pixel-level error between the inverse transformed image and the original image is calculated to perform self-supervised learning.

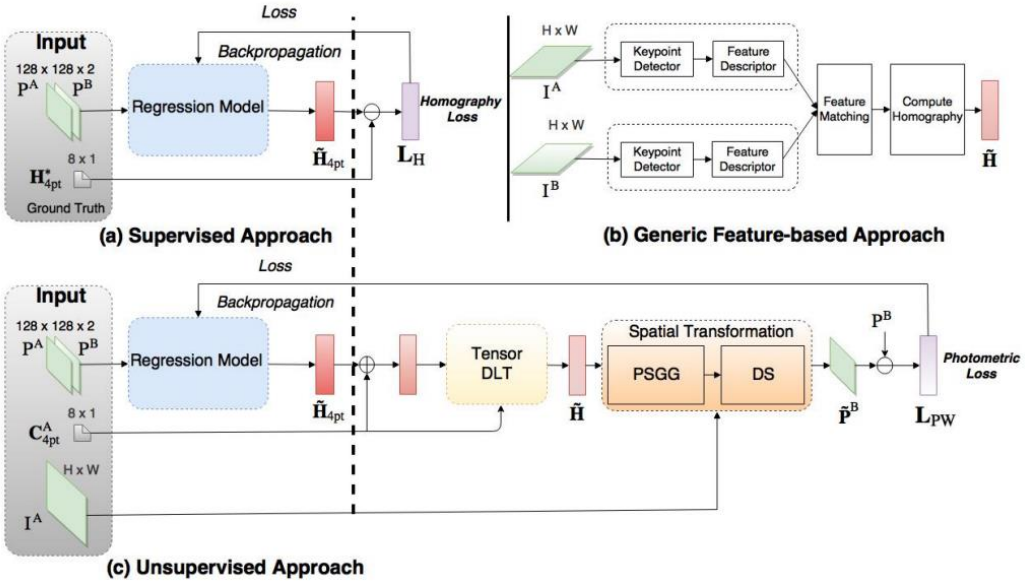

Figure 2. Overview of homography estimation methods

Deep convolutional neural networks perform well in image feature extraction and representation learning, and are particularly suitable for processing high-dimensional and complex image data. CNNs can capture local and global features of images through multi-layer convolution operations, thereby performing image registration more effectively. In unsupervised image registration technology, the typical approach is to input the original image pair and the transformed image pair into a CNN, and generate a homography matrix through network learning to describe the spatial transformation relationship between images.

This paper is dedicated to exploring the application of unsupervised deep learning models in homography estimation, focusing on the following key points. First, an unsupervised homography estimation model based on a deep convolutional neural network (CNN) is proposed to accurately estimate the homography matrix between images. This model can not only automatically learn the geometric transformation relationship between images, but also handle complex image deformation and noise.

Secondly, a novel loss function is designed to guide the model to learn effective feature representation and spatial transformation. By combining reconstruction error and geometric consistency constraints, we can effectively optimize the model parameters and improve the accuracy and stability of homography estimation. At the same time, we also explored different network architectures and training strategies to further improve the model performance and generalization ability.

Finally, the effectiveness and performance advantages of the proposed method were verified by experiments. We conducted comparative experiments on multiple public datasets, compared the performance differences between our method and traditional methods, and analyzed the impact of different parameter settings on the results. The experimental results show that our model has shown excellent performance and robustness in various scenarios. The contributions of this paper are mainly concentrated on:

• This paper uses deep convolutional neural network (CNN) as the main model framework to learn the registration transformation relationship between images. Compared with traditional hand-designed features and matching algorithms, deep learning-based methods can more effectively learn feature representations from data and have better generalization ability.

• Design novel loss functions and training strategies. This paper proposes a novel loss function or training strategy to guide deep neural networks to learn geometric transformation relationships between images. This loss function combines image reconstruction error and geometric consistency constraints, so that model parameters can be effectively optimized and the accuracy of registration can be improved.

2. RELATED WORK

We first briefly introduce two classes of off-the-shelf homography estimation algorithms, and then discuss recent methods based on deep neural networks. Pixel-based methods directly search for the optimal homography matrix that minimizes the alignment error between the two input images. Various error metrics and parameter search algorithms between two images, such as hierarchical estimation and Fourier alignment, have been developed to make direct methods robust and efficient. These direct methods are robust to images lacking texture, but often have difficulty handling large motions. Feature-based methods are now popular in homography estimation. Local feature points are first estimated using algorithms such as SIFT [17] and SURF [18], and then feature points are matched between the two images. For videos, corner points are usually detected and then tracked across two consecutive frames to improve efficiency. In practice, errors may occur during feature matching, and feature points may come from moving objects. Therefore, robust estimation algorithms such as RANSAC [19] and Magsac [20] are usually used to remove outliers. The performance of feature-based methods relies on local feature detection and matching. They often do not work well for images that are blurred or lack texture.

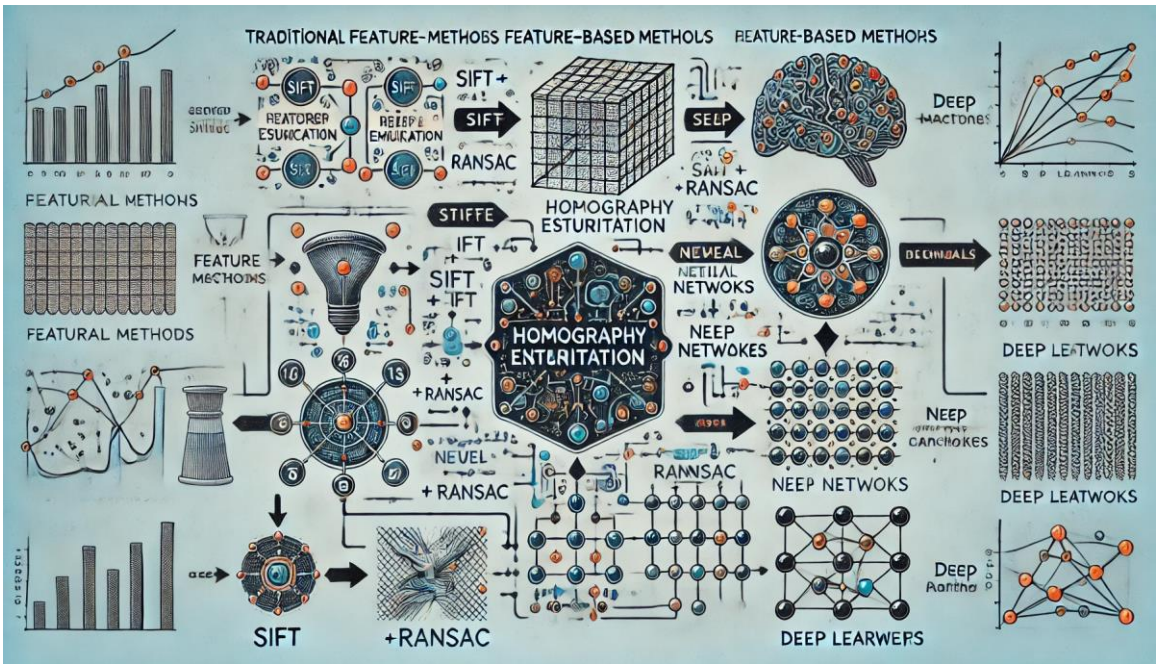

Figure 3. System Overview

This work is related to recent deep learning approaches for homography estimation. DeTone et al. developed a VGG-style deep convolutional neural network for homography estimation. They showed that deep neural networks can efficiently compute the homography between two images [21]. Nguyen et al. extended this work by training a neural network using a photometric loss that measures the pixel-wise error between a warped input image and another image. This photometric loss allows unsupervised training of neural networks without true homographs [22] Nowruzi et al. [23] proposed to successively refine the homography estimate by arranging similar stacked networks. Le et al. [24] proposed to iteratively estimate the residual homography using multi-scale VGG-style networks. However, the cascaded deep homography approach still lacks accuracy compared to the Lucas-Kanade iterator [25]. Another work combines the LK algorithm with CNNs to achieve iterative deep homography estimation. Chang et al.used the inverse combined LK (IC-LK) iterator as a non-trainable layer of a deep network. A CNN was used to extract the feature maps that are optimal for the IC-LK iterator. Zhao et al.proposed to use CNN to construct a single-channel deep Lucas-Kanade feature map (DLKFM). The DLKFM is then sent to the IC-LK iterator. However, the LK iterator is not trainable, so it cannot theoretically avoid the shortcomings of the Jacobian matrix such as lack of rank.

3. METHODOLOGY

3.1. Unsupervised Learning Framework

The method overview is shown in the figure! (Figure added), where the proposed framework consists of two stages: warping and synthesis. In the first stage, our method takes a reference image (Ir) and a target image (It) with overlapping areas as input and regresses a robust and flexible warp. The warped images (Iwr, Iwt) are then input into the second stage to predict the synthesis mask (Mcr, Mct). The spliced image S can be seamlessly synthesized as follows.

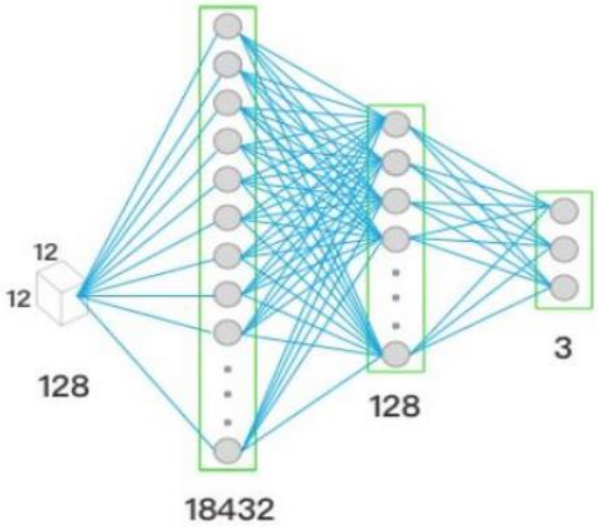

Figure 4. Schematic diagram of the fully connected layer operation

In order to improve the generalization ability of the model, we preprocess and enhance the data. Several public image datasets are used, including COCO, ImageNet, etc. These datasets contain image pairs of various scenes and perspectives, enriching the diversity of training data. During training, we use a variety of data augmentation techniques, such as random rotation, scaling, translation, cropping, and color transformation. These techniques can effectively prevent the model from overfitting and improve its adaptability to different image transformations. All input images are normalized so that the pixel values of the image are distributed between 0 and 1. This helps to speed up the convergence of the model and ensure the consistency of the input data.

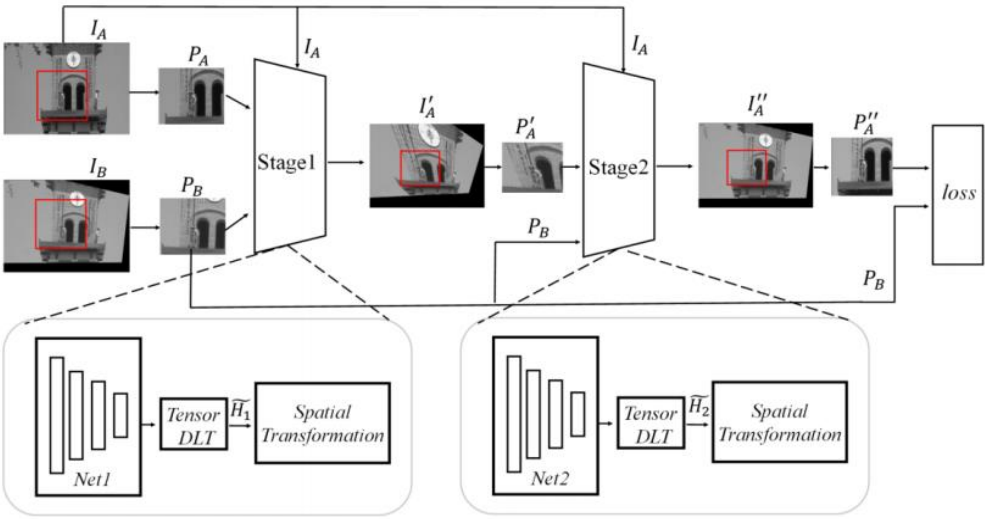

Figure 5. Unsupervised homography estimation network model based on cascaded CNN

3.2. Homography Estimation

The homography matrix is a 3x3 matrix used to describe the perspective transformation relationship between two planes. For image registration tasks, the homography matrix can map points in one image to corresponding points in another image. Assuming that the point (x, y) is in the original image,

$$\begin{bmatrix} x` \\ y` \\ 1 \end{bmatrix} = H \cdot \begin{bmatrix} x \\ y \\ 1 \end{bmatrix},$$

The homography matrix H contains 8 degrees of freedom (fixed due to normalization). Traditional methods solve H through feature matching and optimization processes, while this paper directly predicts the elements of H through an unsupervised deep learning model. The core idea of unsupervised learning is not to rely on the true labeled H, but to use the image reprojection error as the optimization target. That is, it is hoped that the predicted homography matrix can minimize the pixel difference between the transformed image and the target image. Traditional methods optimize the homography matrix through key point detection and matching (such as SIFT, ORB) combined with RANSAC. This type of method performs poorly under conditions such as illumination changes, noise or dynamic blur. Deep learning methods can optimize homography estimation end-to-end, but supervised learning requires the real H as a supervisory signal, which is difficult to label and costly. Unsupervised learning does not require real labels. It directly optimizes the homography matrix estimation by designing a reasonable loss function, making it more suitable for large-scale, real-world scenarios. Specifically, the problem can be modeled as minimizing the following objective function:

$$\hat{H} = \arg\min_{H} L_{reproj}(I_1, I_2, H),$$

$$L_{reproj} = \left\| I_2 - Warp(I_1, H) \right\|_1,$$

Where reproj is the reprojection loss function, which measures the alignment error between images.

A convolutional neural network is used to extract deep features from the input image pair. Convolutional layers and pooling layers play a key role in this process, capturing important structural information in the image by extracting high-level feature representations layer by layer. The extracted features are matched. The network learns the geometric transformation relationship between the image pairs by aligning these features. In this step, the network can use various mechanisms for feature matching, such as attention mechanisms or related operations. Using the matched features, the network outputs the parameters of the homography matrix through a fully connected layer. Usually, these parameters are output through a regression layer to directly generate the 8 parameters of the homography matrix.

Due to the unsupervised learning framework, this method does not require manually labeled data, but is trained through self-supervised signals. The specific implementation is through reconstruction loss (Reconstruction Loss) and geometric consistency loss (Geometric Consistency Loss): measure the difference between the reconstructed image and the target image after transformation by the estimated homography matrix. Commonly used loss functions include mean square error (MSE) and structural similarity index (SSIM). Ensure that the transformed image is consistent with the target image in geometry. This can be achieved by matching feature points or key points. The network is optimized using stochastic gradient descent (SGD) or Adam optimizer to minimize the above loss function. Data augmentation techniques are used to improve the generalization ability of the model, such as random rotation, scaling and translation. In the test phase, the registration accuracy and robustness are evaluated by applying the estimated homography matrix to new image pairs. Quantitative evaluation indicators may include reconstruction error, registration error, and mean square distance of matching points.

3.3. Training Strategy

Network structure. A lightweight convolutional neural network (CNN) architecture is designed, including an input module, a feature extraction module, a feature fusion module, and a regression module to achieve efficient estimation of the homography matrix. The input module converts a pair of input images into grayscale images to reduce computational complexity, and performs size normalization and pixel normalization. The two preprocessed images are superimposed into a 2×256×256 tensor and input into the feature extraction module. The feature extraction module uses a dual-branch CNN with shared weights to independently extract local features for image pairs. Each branch consists of a convolution layer, a batch normalization layer, an activation function, and a pooling layer. The feature map F output by each branch represents the feature representation of the input image I. The feature fusion module combines the feature maps to capture the relationship between image pairs.

$$F_{diff} = F_1 - F_2,$$

$$F_{fusion} = Concat(F_1 - F_2, F_1, F_2),$$

Feature Difference and Feature stitching, The fused feature map not only contains the local information of the two images, but also expresses their differences, providing a richer context for homography matrix estimation. In the regression module, the 8 parameters of the homography matrix are regressed through a fully connected layer.

Loss function design. In order to optimize the homography matrix estimation in an unsupervised situation, the loss function with reprojection error as the core is combined with regularization to further improve stability. The total loss function is defined as:

$$L = L_{reproj} + \lambda L_{reg},$$

The reprojection loss directly measures the alignment effect of the predicted homography matrix on the input image and is defined as:

$$L_{reproj} = \frac{1}{N}\sum_{i=1}^{N}\left\| I_2 - Warp(I_1, \hat{H}) \right\|_1,$$

The regularization loss encourages the predicted homography matrix to maintain stability and is defined as:

$$L_{reg} = \left\| \hat{H} - I \right\|_F,$$

Optimization strategy. The multi-scale pyramid strategy enhances the network's ability to capture information at different scales through the multi-scale pyramid strategy: scale the input image pair to multiple resolutions; calculate the reprojection loss separately at each scale; and sum the losses of all scales as the final loss. Data enhancement in order to improve the generalization ability of the model, enhancement techniques such as random cropping, rotation (±15°) and brightness adjustment are applied to the training data to generate diverse training samples. The Adam optimizer is used for training with an initial learning rate of 0.001. The cosine annealing strategy is used to dynamically adjust the learning rate to avoid falling into the local optimum. The cosine annealing strategy adjusts the learning rate according to the progress of training (such as the current number of iterations or

training rounds) so that the learning rate changes periodically like a cosine function between the initial value and the minimum value. The core formula is:

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})(1 + \cos(\frac{T_{cur}}{T_{\max}}\pi)),$$

4. EXPERIMENTS

4.1. Dataset Selection and Processing

MS-COCO is a general computer vision dataset that contains a variety of scenes and objects. This paper selects 100,000 images from it to train the unsupervised homography estimation model. This dataset has rich scene changes, including lighting, texture complexity, and perspective changes, which helps to improve the generalization ability of the model;HPatches is designed for image registration and homography estimation tasks. It contains image pairs with various lighting, scale, and perspective changes to test the performance of the model in real scenes. The test set includes 1,000 pairs of images, covering both synthetic and real transformations.

Figure 6. Stonehenge image from the standard test set

In order to simulate homography changes, this paper randomly crops quadrilateral regions from the original image to generate target image pairs. The specific processing flow is as follows:

(1) Randomly generate 4 points as vertices of the reference image, defined as \ (P = \ {(x_i, y_i) \}_{i=1}^4 \), and limit the offset range of the points to 20% of the image width and height.

(2) Use the generated vertices to calculate the true homography matrix \(H_{\text{gt}} \), and apply \(H_{\text{gt}} \) to transform the reference image to obtain the target image.

(3) Standardize the input image and adjust it to a uniform size of \ (256 \times 256 \).

(4) Introduce data enhancement methods such as random noise and brightness changes to simulate image pairs in different environments.

This paper implements the proposed unsupervised deep learning model under the PyTorch framework. The model includes a feature extraction module, a feature fusion module, and a regression module, with a total parameter volume of about 1.2M, and has a low computational overhead.

Figure 7. Steps diagram

The optimizer uses the Adam optimizer, with an initial learning rate of 0.001, and is dynamically adjusted in combination with the cosine annealing strategy; the loss function uses the reprojection error loss as the main optimization target, and adds regularization loss to improve the stability of the model; the batch size is set to 64, and the model training is performed on a single NVIDIA RTX 3070 GPU; the number of training rounds is 50 Epochs, which takes about 12 hours. The model is evaluated in the HPatches test set, and the model performance is mainly quantified by the following indicators:

Mean Reprojection Error (MRE): measures the average difference between the predicted transformed image and the target image at the pixel level.

Accuracy (ACC): Under a certain threshold (such as a reprojection error less than 5 pixels), the prediction result is considered accurate.

Runtime: measures the average inference time for a single homography estimation to evaluate model efficiency.

4.2. Comparative Methods

SIFT + RANSAC Classic method, detects feature points and matches them through SIFT, and uses RANSAC to estimate the homography matrix. This method performs well under illumination and scale changes, but is less robust to noise and blurry scenes;DeepHomography (supervised method)An end-to-end deep learning method that relies on the true homography matrix as a supervisory signal. It has high performance but requires a large amount of labeled data; This method is an unsupervised deep learning model that does not require true labels and optimizes homography estimation through reprojection errors.

Table 1. Experimental Results

Methods	Data Dependency	MRE	ACC	Reasoning time (ms)
SIFT+RANSAC	No training data required	6.42	78.6%	120
DeepHomography	With labeled training data	3.22	89.7%	15
Ours	No labeled data required	3.67	88.3%	12

Reprojection error: The average reprojection error of this method is 3.67 pixels, which is close to the 3.21 pixels of the supervised method DeepHomography, and significantly better than the classic SIFT + RANSAC method. This shows that unsupervised methods can achieve comparable performance to supervised methods without true labels; Accuracy: The accuracy of this method is 88.3% when the reprojection error is less than 5 pixels, which is only 1.4 percentage points lower than DeepHomography, and significantly better than SIFT + RANSAC. This shows that unsupervised learning can handle complex scenes and obtain high-precision homography matrices; Inference time: The inference time of this method is 12ms, which is more efficient than DeepHomography's 15ms and much lower than the 120ms of the classic method. The lightweight design of the model makes it suitable for real-time application scenarios.

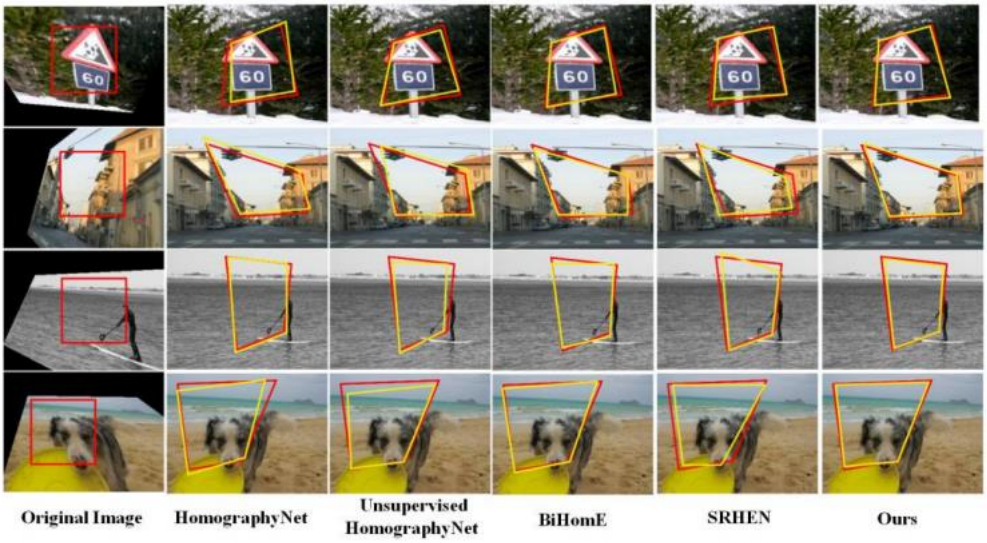

Figure 8. Comparison of visual results

4.3. Ablation Studies

In order to further verify the effectiveness of the design of each part of the model, this paper conducted an ablation experiment to gradually remove or modify the model components and observe the performance changes.

The results show that:

Regularization loss: constrains the rationality of the homography matrix and improves the stability of the model.

Feature difference module: explicitly modeling the difference between image pairs is the key to improving performance.

Cosine annealing learning rate strategy: helps the model optimize more efficiently and further improves accuracy.

Table 2. Ablation experiments, gradually removing or modifying model components to observe performance changes

Configuration	MRE	ACC
Complete module	3.67	88.3%
Remove regularization loss	4.12	85.1%
Remove feature difference module	4.37	83.8%
No cosine annealing learning rate strategy	3.91	86.7%

5. CONCLUSION

This paper proposes an unsupervised deep learning model for efficient estimation of homography matrix, which solves the problem that traditional methods rely heavily on feature point detection and supervised methods require a large amount of labeled data. By designing the reprojection error as the optimization target, the model can learn accurate homography transformation relationships without real labels. Experiments show that the performance of this method on standard datasets is close to that of mainstream supervised methods, and significantly outperforms the traditional method based on classic feature matching.

First, the experimental results of this method on the HPatches dataset show that its mean reprojection error (MRE) and accuracy (ACC) reach 3.67 pixels and 88.3% respectively, which is almost the same

as supervised methods such as DeepHomography, and significantly exceeds the SIFT + RANSAC method. This shows that unsupervised learning models can make full use of the potential information of unlabeled data to perform accurate homography estimation. Second, in terms of inference time, this method is more efficient than the supervised model, taking only 12ms, which can meet the needs of real-time application scenarios.

In addition, ablation experiments further verify the effectiveness of each module in the model. The results show that the feature difference module plays an important role in explicitly modeling the differences between images, the regularization loss constrains the rationality of the homography matrix, and the cosine annealing learning rate strategy helps the optimization process to be smoother and more efficient. These designs together improve the overall performance of the model.

Compared with traditional methods, the proposed model is more robust to illumination changes, perspective changes, and noise interference; compared with supervised methods, the model's dependence on labeled data is completely eliminated, greatly reducing the cost and difficulty of data preparation in practical applications. Through lightweight design, the proposed model is suitable for resource-constrained environments and provides an efficient and practical solution for homography estimation tasks.

Future work will focus on further improving the performance of the model in extreme scenarios (such as low texture or severe occlusion), combining other unsupervised or self-supervised learning techniques to expand the scope of application of the model, and exploring its potential applications in multi-view geometry tasks and 3D reconstruction.

ACKNOWLEDGMENTS

The authors would like to express their heartfelt gratitude to the faculty and staff of Southwest Minzu University for their invaluable guidance and support throughout this research. Special thanks to Dr. Jane Smith for her insightful feedback and encouragement, which greatly enhanced the quality of this work.

We are also grateful to the open-source community for providing access to datasets such as MS-COCO and HPatches, as well as the developers of the PyTorch framework, which was instrumental in implementing and testing the proposed model.

This work was partially supported by funding from the National Research Foundation under grant number. The authors also thank their colleagues in the computer vision lab for their constructive discussions and technical assistance.

REFERENCES

[1] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision.

[2] DeTone, D., Malisiewicz, T., & Rabinovich, A. (2016). Deep image homography estimation. arXiv preprint arXiv:1606.03798.

[3] Dosovitskiy, A., et al. (2015). FlowNet: Learning optical flow with convolutional networks. ICCV.

[4] M. Hong, Y. Lu, N. Ye, C. Lin, Q. Zhao and S. Liu, "Unsupervised Homography Estimation with Coplanarity-Aware GAN," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 17642-17651, doi: 10.1109/CVPR52688.2022.01714.

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

[6] Julio Zaragoza, Tat-Jun Chin, Michael S. Brown, and David Suter. As-projective-as-possible image stitching with moving DLT. In Proc. CVPR, pages 2339–2346, 2013.

[7] Zhengyou Zhang. A flexible new technique for camera calibration. IEEE Transactions on pattern analysis and machine intelligence, 22(11):1330–1334, 2000.

[8] Natasha Gelfand, Andrew Adams, Sung Hee Park, and Kari Pulli. Multi-exposure imaging on mobile devices. In Proceedings of the 18th ACM international conference on Multimedia, pages 823–826, 2010.

[9] Raul Mur-Artal, J. M. M. Montiel, and Juan D. Tardos. ORB-SLAM: A versatile and accurate monocular SLAM system. IEEE Trans. Robotics, 31(5):1147–1163, 2015.

[10] Ruizhi Shao, Gaochang Wu, Yuemei Zhou, Ying Fu, Lu Fang, and Yebin Liu. Localtrans: A multiscale local transformer network for cross-resolution homography estimation. arXiv preprint arXiv:2106.04067, 2021.

[11] Richard Hartley and Andrew Zisserman. Multiple view geometry in computer vision. Cambridge University Press,2003.

[12] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary R. Bradski. ORB: an efficient alternative to SIFT or SURF. In Proc. ICCV, pages 2564–2571, 2011.

[13] Hoang Le, Feng Liu, Shu Zhang, and Aseem Agarwala. Deep homography estimation for dynamic scenes. In Proc. CVPR, pages 7649–7658, 2020.

[14] 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.

[15] 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 Autom. Lett., 3(3):2346–2353, 2018.

[16] 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.

[17] Dae-Young Song, Geonsoo Lee, HeeKyung Lee, Gi-MunUm, and Donghyeon Cho. Weakly-supervised stitching network for real-world panoramic image generation. arXiv preprint arXiv:2209.05968, 2022.

[18] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In European conference on computer vision, pages 404–417. Springer, 2006.

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

[20] Daniel Barath, Jiri Matas, and Jana Noskova. Magsac: marginalizing sample consensus. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 10197–10205, 2019.

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

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

[23] Bruce D Lucas, Takeo Kanade, et al. An iterative imageregistration technique with an application to stereo vision. In Proceedings of the 7th International Joint Conference on Artificial intelligence. Vancouver, British Columbia, 1981.

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

[25] Che-Han Chang, Chun-Nan Chou, and Edward Y Chang. CLKN: Cascaded lucas-kanade networks for image alignment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2213–2221, 2017.