ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446158

DEPTH-GUIDED DOMINANT PLANE PERCEPTION FOR UNSUPERVISED HOMOGRAPHY ESTIMATION

Xiaomei Feng [1], *Qi Jia* [1*], *Yu Liu* [1*], *Xin Fan* [1], *Longin Jan Latecki* [2]

[1] Dalian University of Technology [2] Temple University

ABSTRACT

Homography describes the mapping relations of the same plane across views. In scenarios with multiple planes, single homography estimation aims to obtain the optimal solution generated by the largest consistent plane to obey the coplanar constraints. However, existing methods typically consider all planes equally, neglecting the negative impact of regions that differ significantly from the largest approximate planar areas (dominant plane). In this work, we propose a depth-guided dominant plane perception network to achieve unsupervised homography estimation with additional attention on the dominant plane. Specifically, we leverage the depth-wise prior to adaptively detecting the approximate dominant plane, invoking essential scene structures for unsupervised homography estimation. Then, we enhance the corresponding features of the dominant plane and explore their correlations through a specially designed perceptual module. Finally, we employ dominant plane perception on multi-scale features progressively to estimate the homography in a coarse-to-fine manner. Extensive experiments on a large parallax dataset demonstrate that our method improves the alignment performance by 10.29%, yielding more accurate alignment than previous competitive methods.

Index Terms— Homography estimation, depth guidance, image alignment

1. INTRODUCTION

Homography describes the mapping relations of the same plane across views, and it is widely used in multi-view image processing, such as image/video stitching [1, 2], image alignment [3, 4], and SLAM [5]. In the scenarios with multiple planes, the homography estimated by the approximate maximal planar regions can be regarded as the optimal solution, that aligns significantly large areas as well as minimizes the alignment error on the other planes [6].

To estimate the approximate optimal solution, most of the existing methods capture more matching features to cover each plane. For example, traditional methods [7–10] introduce extra structures to generate more valid matching features, and they also employ the outlier removal strategy [11–13] to select consistent features. Unfortunately, these methods still suffer from the lack of discriminative features under low-texture or blurry image pairs. In contrast, deep learning-based methods [14–18] capture dense matching features. However, these methods consider all matching features equally and ignore the negative impact of features that differ significantly from the large consistent planes, leading to inaccurate alignments in the dominant plane, as illustrated in Fig. 1.

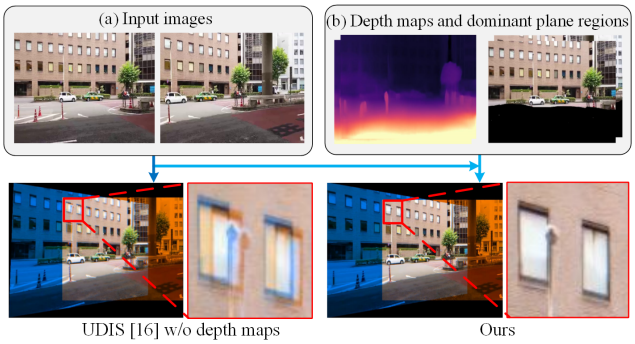

Fig. 1: Visualization of image alignment results with and without depth guidance. The result of UDIS [16] w/o depth maps employs only (a) as input, and our result leverages both (a) and (b) as inputs. The alignment results are generated by superimposing the warped target and reference images, and the misaligned pixels are visualized as colored ghosts.

To focus on the discriminative features of large consistent planes, some existing methods [6, 19, 20] predict a mask to remove large foregrounds or moving objects. However, they predicted mask is a side product of homography estimation, that lacks reliable guidance for scene structure information and easily fails in image pairs with large parallax. The dominant plane is crucial for generating the optimal solution of homography, and we focus on the dominant plane to estimate homography, thus significantly reducing the artifacts and misalignment as illustrated in Fig. 1 the magnified part marked with the red box. Consequently, our main idea is to detect and leverage the discriminative features of the dominant plane for accurate homography estimation.

In this work, we propose a depth-guided dominant plane perception network to generate the approximate optimal solution for unsupervised homography estimation. Specifically, we leverage the prior knowledge of depth maps to divide multiple planes and select the plane with the largest regions as a dominant plane. Moreover, we develop a dominant plane perception module that guides the network to focus on the dominant planes and explore their correlation over multi-scale features. Finally, we employ both global and local alignment loss to promote content alignment and strengthen the alignment of the dominant plane. To summarize, our main contributions are threefold: (1) We invoke depth-guided scene decomposition to select the plane with maximal consistency for addressing the homography estimation of scenes with multiple planes. (2) We design a dominant plane perception module to enhance the features of the dominant plane and explore their correspondences to estimate an approximately optimal homography without ground truth. (3) Our method achieves state-of-the-art performance and outperforms previous methods by 10.29% for alignment on real-world datasets.

*equal corresponding author. This work was supported in part by the NSF of China under Grant Nos.62272083, 62102061 and U22B2052, and in part by the Liaoning Provincial NSF under Grant 2022-MS-137 and 2022-MS-128.

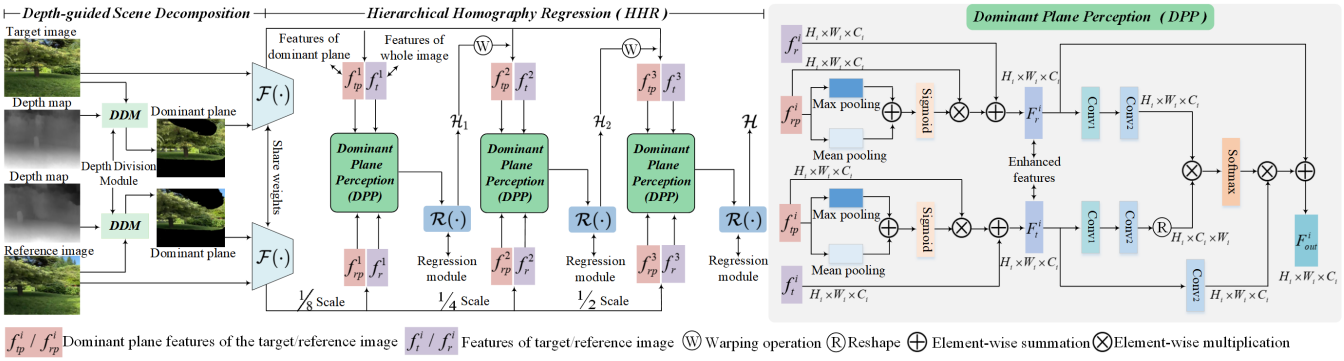

Fig. 2: The overall pipeline of the proposed method. The image pairs with parallax and their correspondence depth map as input, and the depth-guided scene decomposition uses the depth division module (*DDM*) to detect the dominant plane. We design the dominant plane perception module (*DPP*) to enhance the features of the domain plane, and it is inserted into multi-scale features to regress final homography $\mathcal{H}$ by hierarchical homography regression (*HHR*).

2. METHOD

Overview: Figure 2 demonstrates the overall structure of the proposed unsupervised homography estimation network with depth-guided dominant plane perception learning. Given a pair of target image I_t and reference image I_r, we obtain their corresponding depth maps I_t^d and I_r^d by existing depth estimation network [21]. Firstly, we leverage the depth-guided scene decomposition with the depth division module *DDM* to select the largest approximately planar regions P_t^m and P_r^m (Sec. 2.1). Then, we design a dominant plane perception module (*DPP*) to enhance features of the dominant plane at different feature scales (Sec. 2.2). Based on the enhanced features, we employ a hierarchical homography regression (*HHR*) mechanism, starting from the 1/8 feature scale at level 1 and finishing with the 1/2 feature scale at level 3 (Sec. 2.3). Finally, we optimize the model with the unsupervised objective function (Sec. 2.4).

2.1. Depth-guided Scene Decomposition

We utilize the depth prior to dividing the depth map into different regions based on the depth value of each pixel. The depth regions are used as masks to select corresponding approximate plane regions in the RGB image. Firstly, we employ the pre-trained model [21] to predict raw depth maps I_r^d and I_t^d. Then, as shown in the depth division module *DDM* of Fig. 3 (a), we quantize the raw depth map by histogram and select the first N largest modes of the multimodal depth distributions to decompose the original depth map into $N+1$ regions $D_1, D_2, ...D_{N+1}$, including N depth interval windows and a window of the remaining part. Finally, we generate $N+1$ binary masks $M^j, j \in [1, 2, ...N+1]$, where each mask corresponds to a region, and pixels within a region have a pixel value of 1. Consequently, we can divide the original image into $N+1$ approximate planes by:

$$P_r^j = I_r \otimes M_r^j, P_t^j = I_t \otimes M_t^j, j \in [1, 2, ...N+1], \quad (1)$$

where P_r^j and P_t^j represent the j-th plane in reference and target image respectively, $\otimes$ represents the element-wise multiplication. M_t^j and M_r^j represent the masks of target and reference images, respectively. Then, we select the plane with the largest areas $P_r^m = \max(P_r^j)$, $P_t^m = \max(P_t^j)$ as the dominant plane.

As demonstrated in Fig. 3 (b), we decompose the depth maps and generate the corresponding planar regions P^1, P^2, and P^3 with similar depth. In this work, we set $N = 2$ to decompose the original depth map into three regions, and our ablation experiments validate the performance of different depth map division numbers.

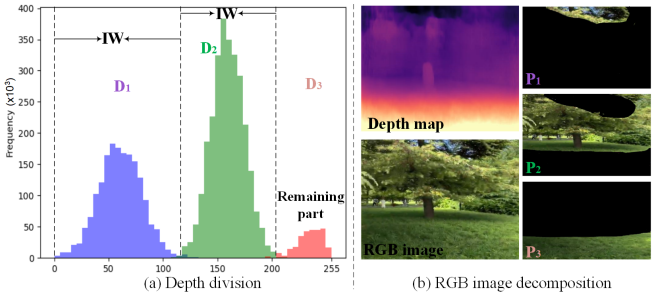

Fig. 3: Schematic of depth division and RGB image decomposition. 'IW' represents the depth interval window, and D_1, D_2, and D_3 represent different depth division regions. P_1, P_2, and P_3 indicate the decomposed different planar regions with similar depth values.

2.2. Dominant Plane Perception

Based on the detected dominant plane, we design a dominant plane perception module *DPP* that guides the network to focus on the dominant plane in multi-scale features, as shown in the right part of Fig. 2. Specifically, we fuse the features of the entire image and the dominant plane to enhance the representation of dominant plane features and then leverage correlation calculations to capture discriminative features for homography estimation.

Firstly, we apply a max-pooling layer and an average-pooling layer in parallel to remove noise and preserve significant features for dominant plane images f_{rp}^i and f_{tp}^i. After a sigmoid layer, the feature of the dominant plane enhances the entire image feature by:

$$F_r^i = f_{rp}^i \otimes S(max(f_{rp}^i) \oplus avg(f_{rp}^i)) \oplus f_r^i, i \in [1, 2, 3], \quad (2)$$

where F_r^i represents the enhanced feature of the reference image on the i-th feature scale. S indicates a sigmoid layer, and $\oplus$ indicates element-wise summation. The same operations are applied on f_{tp}^i and f_t^i to generate enhanced feature F_t^i.

Subsequently, we explore the correlation between the enhanced features F_t^i and F_r^i. We leverage two convolutional layers $Conv_1$ and $Conv_2$ with the 3×3 and the 1×1 convolutional kernels to activate features and employ matrix multiplication to calculate the

correlation C^i in each feature scale by:

$$C^i = (Conv_2(Conv_1(F_r^i))) \otimes (\circledR(Conv_2(Conv_1(F_t^i)))), \quad (3)$$

where $i \in [1, 2, 3]$, and $\circledR$ indicates reshape operation.

To focus on discriminative features, we leverage a softmax layer $\mathcal{S}$ to normalize the correlation map and apply an element-wise multiplication on activated F_t^i. Finally, the final dominant plane perception feature F_{out}^i is obtained as:

$$F_{out}^i = Conv_2(F_t^i) \otimes \mathcal{S}(C^i) \oplus F_r^i, i \in [1, 2, 3]. \quad (4)$$

2.3. Hierarchical Homography Regression

To consider features in both shallow and deep feature scales simultaneously, we utilize a hierarchical mechanism with a recurrent refinement strategy to estimate offsets of four image vertexes, which are then combined with direct linear transformation (DLT) [2] to generate the homography matrix. Specifically, we initially leverage the dominant plane perceptual features in 1/8 scale to regress offsets Δ_1 and corresponding $\mathcal{H}_1$ through regression module $\mathcal{R}(\cdot)$. The module $\mathcal{R}(\cdot)$ is composed of four Conv+BN+ReLU convolutional blocks and a linear block. Then $\mathcal{H}_1$ performs a coarse alignment for 1/4 feature scale of the target image to estimate residual offsets Δ_2. After that, a similar operation is used for 1/2 feature scale to estimate residual offsets Δ_3. In hierarchical homography regression, we accumulate the residual offsets of each scale to estimate the final homography $\mathcal{H}$:

$$\mathcal{H} = DLT(\sum_{i=1}^{3} \Delta_i), i \in [1, 2, 3]. \quad (5)$$

2.4. Unsupervised Training

Our total loss function $\mathcal{L}_{total}$ consists of $\mathcal{L}_{global}$ and $\mathcal{L}_{local}$, expressed as Eq. 6. $\mathcal{L}_{global}$ is the global alignment loss to ensure the alignment of the entire image, and meanwhile $\mathcal{L}_{local}$ is a local alignment loss to align the dominant planes.

$$\mathcal{L}_{total} = \lambda \mathcal{L}_{global} + (1 - \lambda)\mathcal{L}_{local}, \quad (6)$$

where λ is a trade-off parameter of two terms. For each scale, we impose unsupervised constraints to minimize L_1 distance between the warped target image and its corresponding reference image by:

$$\mathcal{L}_{global} = \sum_{i=1}^{3} \left\| \mathcal{H}_i(I_t) - \mathcal{H}_i(E) \otimes I_r \right\|_1, \quad (7)$$

where $\mathcal{H}_i$ is the predicted homography at i-th feature scale, and E is an all-one matrix with identical size with I_r. In addition, we also leverage the same loss to constrain the alignment of the dominant planes P_t^m and P_r^m by $\mathcal{L}_{local}$.

3. EXPERIMENTS

Dataset: We evaluate our method on the most widely used UDIS-D dataset [16] to keep the same setting with [16] and [17], which contains 10440 training images and 1106 testing images with a size of 512×512. UDIS-D is a real-world dataset that includes a wide variety of challenging image pairs with different overlap rates and parallax such as indoor, nighttime, dark, and snowy.

Details: During the training phase, we resize each image to 128×128 as input. We set $N = 2$ for depth map division and the trade-off parameter in the loss function is $\lambda = 0.9$. Our network is implemented in PyTorch [22] and runs an NVIDIA RTX 3090 GPU. The learning rate is 1×10^{-4} and the training process consists of 150 epochs with a batch size of 64.

3.1. Comparison with Existing Methods

We compare our method with two categories of existing homography estimation methods: (1) Deep learning-based methods including DHN [14], UDHN [15], CA-UDHN [19], UDIS [16], and MGDH [17]. (2) Traditional methods including APAP [2], ELA [7], SPW [8], and LPC [9]. We employ PSNR [23] and SSIM [24] metrics to measure the performance of different methods.

Quantitative comparison: Tables 1 and 2 demonstrate the quantitative comparison results of our method with deep learning-based methods and traditional methods, respectively. We divide the test results into three levels according to their performance, including 'Easy (Top 0-30%)', 'Moderate (Top 30-60%)', and 'Hard (Top 60-100%)'. For fairness, we use the same image resolution for all inputs to evaluate the performance.

Table 1: Quantitative comparison with cutting-edge deep learning based methods on UDIS-D dataset with 128×128 resolutions. Bold indicates the best performance.

Methods	PSNR ↑				SSIM ↑			
	Easy	Moderate	Hard	Average	Easy	Moderate	Hard	Average
DHN [14]	16.40	13.36	11.48	13.52	0.409	0.170	0.076	0.204
UDHN [15]	19.39	15.93	13.09	15.83	0.573	0.334	0.165	0.338
CA-UDHN [19]	18.05	13.13	11.00	13.16	0.339	0.181	0.105	0.198
UDIS [16]	27.84	23.95	20.70	23.80	0.902	0.830	0.685	0.793
MGDH [17]	28.41	24.63	21.59	24.54	0.913	0.853	0.733	0.823
Ours	**30.97**	**26.99**	**22.20**	**26.25**	**0.946**	**0.901**	**0.742**	**0.850**

Table 2: Quantitative comparison with traditional methods and the second-best deep learning-based method on UDIS-D dataset with 512×512 resolutions. Bold indicates the best performance.

Methods	PSNR ↑				SSIM ↑			
	Easy	Moderate	Hard	Average	Easy	Moderate	Hard	Average
APAP [2]	27.96	24.39	18.55	23.27	0.901	0.837	0.682	0.794
ELA [7]	29.36	25.10	19.19	24.01	0.917	0.855	0.691	0.808
SPW [8]	26.98	22.67	16.77	21.60	0.880	0.758	0.490	0.687
LPC [9]	26.94	22.63	19.31	22.59	0.878	0.764	0.610	0.736
MGDH [17]	29.52	25.24	21.20	24.89	0.923	0.859	0.708	0.817
Ours	**30.60**	**26.78**	**22.12**	**26.05**	**0.956**	**0.925**	**0.845**	**0.902**

Table 1 demonstrates that our method achieves state-of-the-art performance on all performance levels and outperforms others by a large margin. Our average PSNR is 26.25 dB, which is 10.29% and 6.97% higher than the next two methods UDIS [16] and MGDH [17], respectively. CA-UDHN [19] shares a similar idea that employs the predicted mask to focus on the dominant plane, but its PSNR is 49.87% lower than that of our method. It indicates that we leverage the prior knowledge of depth maps to detect the more reliably dominant plane, yielding more accurate alignment results.

In Table 2, we utilize 512×512 resolutions to evaluate the performance of traditional methods as these methods typically fail to extract features at 128×128 resolutions. Table 2 demonstrates that our method exhibits tremendous advantages on all metrics, and

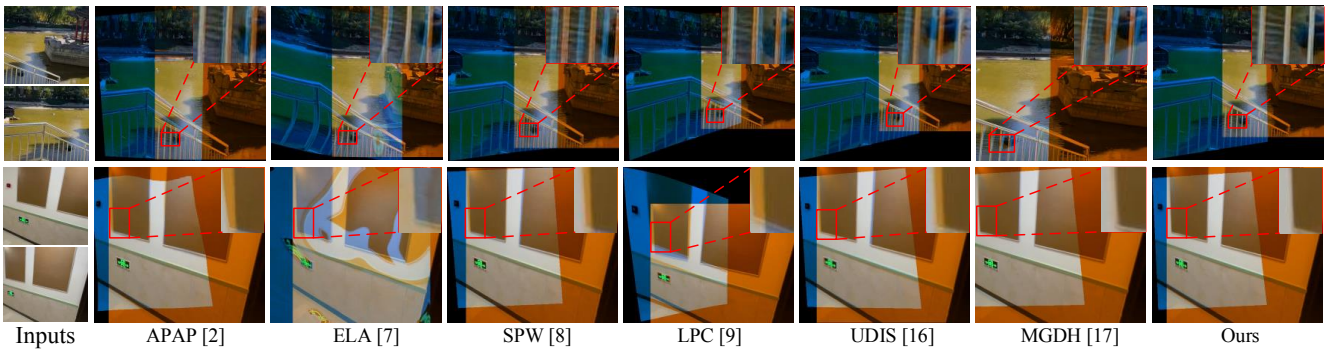

Fig. 4: Qualitative comparison of our method and other cutting-edge methods. The alignment results are generated by superimposing the warped target and reference images, and the highlight regions are overlapping alignment areas, while the non-overlapping regions are visualized as colored ghosts. The red box on the upper right is the zoomed-in region captured in the overlapping regions.

our PSNR and SSIM still outperform the second-best deep learning-based method MGDH. In addition, our PSNR is 2.04 dB higher than the second-best traditional method ELA. Our superior performance validates our dominant plane perception strategy significantly enhances the crucial features compared to the outlier removal strategy in traditional methods.

Table 3: Ablation analysis for the different components on UDIS-D dataset. Bold indicates the best performance.

Metrics	w/o *DPP*	w/o Level 1	w/o Level 1+2	w/o $\mathcal{L}_{local}$	Ours
PSNR ↑	24.94	25.70	24.17	26.15	**26.25**
SSIM ↑	0.806	0.835	0.783	0.847	**0.850**

Table 4: Ablation analysis for approximate plane division on UDIS-D dataset. N indicates the number of depth interval windows and $P = N + 1$ represents the number of correspondence approximate planes. Bold indicates the best performance.

Metrics	$N = 0, P = 1$	$N = 1, P = 2$	$N = 2, P = 3$	$N = 3, P = 4$
PSNR ↑	24.94	26.13	**26.25**	26.19
SSIM ↑	0.806	0.846	**0.850**	0.843

Qualitative comparison: We visualize the alignment results of our method and the six most related comparison methods in Fig. 4. The first row is an outdoor complex scene and the second row is an indoor low-texture scene. As highlighted in the red box and the corresponding zoomed-in regions, our method aligns two images accurately without any artifacts, while other methods exhibit severe deformation and ghosts. ELA fails in the low-textured image as it is difficult to extract discriminative features. The deep learning-based methods (UDIS [16] and MGDH [17]) predict homography based on the entire image, which is easily influenced by features on non-dominant planes. In contrast, our method focuses on the dominant plane to estimate homography, rendering clear and natural alignment results.

3.2. Ablation Studies

We employ a series of ablation studies to validate the effectiveness of the proposed method including different components and the number of plane divisions.

The effectiveness of different components: To validate the effectiveness of the *DPP* module, $\mathcal{L}_{local}$ loss, and *HHR* module, we test performance with or without the corresponding components. Table 3 demonstrates that the performance without *DPP* decreases 5.25% on PSNR metric, while the performance without $\mathcal{L}_{local}$ only decreases 0.3% as the global alignment loss also includes the alignment of the enhanced dominant plane. In addition, the accuracy of the alignment results decreases dramatically without Level 1 or Level 1 + 2 of the *HHR* module. Fig. 5 illustrates image alignment results without corresponding components. As demonstrated in the zoomed-in regions, the alignment results suffer from various degrees of misalignment and artifacts. On the contrary, our method involved in all the proposed modules exhibits accurate alignment in the zoomed-in region, validating the effectiveness of the proposed components.

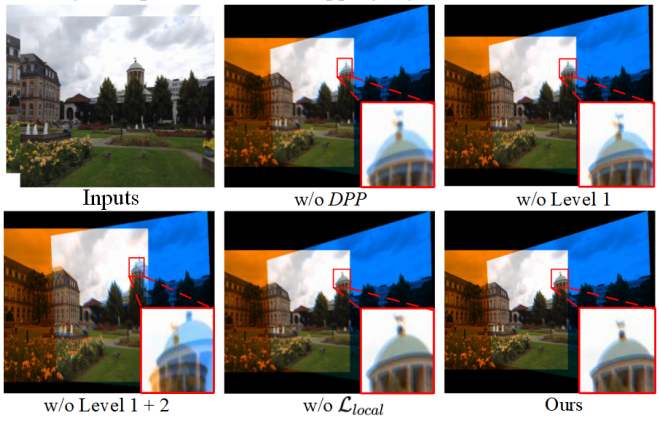

Fig. 5: Visualization of alignment results without corresponding component, where the zoomed-in regions on the bottom right exhibit severe artifacts.

The number of plane divisions: To select the optimal number of approximate plane divisions, we test performance with different N values. Table 4 indicates that our method achieves the best performance when $N = 2, P = 3$. When $N > 2$, the scenarios may be split too finely and affect the performance of alignment.

4. CONCLUSION

We noticed the problem of multiple planes in homography estimation and leveraged a depth-guided dominant plane perception to solve it. Compared with the state-of-the-art methods, our method achieves excellent alignment results on real-world scenes. In the future, we will explore the learning-based dominant plane segmentation method, invoking an adaptive and accurate dominant plane detection for homography estimation.

5. REFERENCES

[1] 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.

[2] Julio Zaragoza, Tat-Jun Chin, Michael S Brown, and David Suter. As-projective-as-possible image stitching with moving dlt. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2339–2346, 2013.

[3] Richard Szeliski et al. Image alignment and stitching: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 2(1):1–104, 2007.

[4] Jiacheng Ying, Hui-Liang Shen, and Si-Yuan Cao. Unaligned hyperspectral image fusion via registration and interpolation modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021.

[5] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

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

[7] Jing Li, Zhengming Wang, Shiming Lai, Yongping Zhai, and Maojun Zhang. Parallax-tolerant image stitching based on robust elastic warping. *IEEE Transactions on multimedia*, 20(7):1672–1687, 2017.

[8] Tianli Liao and Nan Li. Single-perspective warps in natural image stitching. *IEEE transactions on image processing*, 29:724–735, 2019.

[9] Qi Jia, ZhengJun Li, Xin Fan, Haotian Zhao, Shiyu Teng, Xinchen Ye, and Longin Jan Latecki. Leveraging line-point consistence to preserve structures for wide parallax image stitching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12186–12195, 2021.

[10] Peng Du, Jifeng Ning, Jiguang Cui, Shaoli Huang, Xinchao Wang, and Jiaxin Wang. Geometric structure preserving warp for natural image stitching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3688–3696, 2022.

[11] 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.

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

[13] Daniel Barath, Jiri Matas, and Jana Noskova. Magsac: marginalizing sample consensus. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10197–10205, 2019.

[14] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Deep image homography estimation. *arXiv preprint arXiv:1606.03798*, 2016.

[15] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[16] Lang Nie, Chunyu Lin, Kang Liao, Shuaicheng Liu, and Yao Zhao. Unsupervised deep image stitching: Reconstructing stitched features to images. *IEEE Transactions on Image Processing*, 30:6184–6197, 2021.

[17] Lang Nie, Chunyu Lin, Kang Liao, Shuaicheng Liu, and Yao Zhao. Depth-aware multi-grid deep homography estimation with contextual correlation. *arXiv preprint arXiv:2107.02524*, 2021.

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

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

[20] 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 (CVPR)*, June 2020.

[21] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. October 2019.

[22] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[23] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008.

[24] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.