Mask-Homo: Pseudo Plane Mask-Guided Unsupervised Multi-Homography Estimation

Yasi Wang[1], Hong Liu[2], Chao Zhang[1], Lu Xu[1], Qiang Wang[1],

[1]Samsung Research China – Beijing (SRC-B), China
[2]Department of Biomedical Engineering, Eindhoven University of Technology, Netherlands
yasi.wang@samsung.com, h.liu2@tue.nl, {c0502.zhang, lu94.xu, qiang.w}@samsung.com

Abstract

Homography estimation is a fundamental problem in computer vision. Previous works mainly focus on estimating either a single homography, or multiple homographies based on mesh grid division of the image. In practical scenarios, single homography is inadequate and often leads to a compromised result for multiple planes; while mesh grid multi-homography damages the plane distribution of the scene, and does not fully address the restriction to use homography.

In this work, we propose a novel semantics guided multi-homography estimation framework, Mask-Homo, to provide an explicit solution to the multi-plane depth disparity problem. First, a pseudo plane mask generation module is designed to obtain multiple correlated regions that follow the plane distribution of the scene. Then, multiple local homography transformations, each of which aligns a correlated region precisely, are predicted and corresponding warped images are fused to obtain the final result. Furthermore, a new metric, Mask-PSNR, is proposed for more comprehensive evaluation of alignment. Extensive experiments are conducted to verify the effectiveness of the proposed method. Our code is available at *https://github.com/SAITPublic/MaskHomo*.

Introduction

Homography (H) estimation is a fundamental problem in computer vision, that has been extensively used in various applications, such as image alignment, image stitching, etc. A homography is a type of projective transformation, which can be used to describe the mapping relationship between the pixel coordinates of two planes within an image pair.

Traditional H estimation solutions are feature-based, which follow the pipeline of feature detection and matching, outlier rejection and numerical calculation. However, they are highly dependent on the feature detection quality, leading to inaccurate estimation in low texture scenes.

In recent years, unsupervised deep learning-based methods (Nguyen et al. 2018; Ye et al. 2021) which directly predict H by minimizing the difference between the warped source image and target image become top performers. However, an optimal H can be obtained only under the following constraints: *(1) rotation only movements of the camera; (2) the scene locates at a planar surface; (3) the scene is at a distance from the observer*. Therefore, when dealing with scenes containing multiple planes, single-H encounters difficulties in locating and aligning the corresponding regions within the image pair (Fig 1).

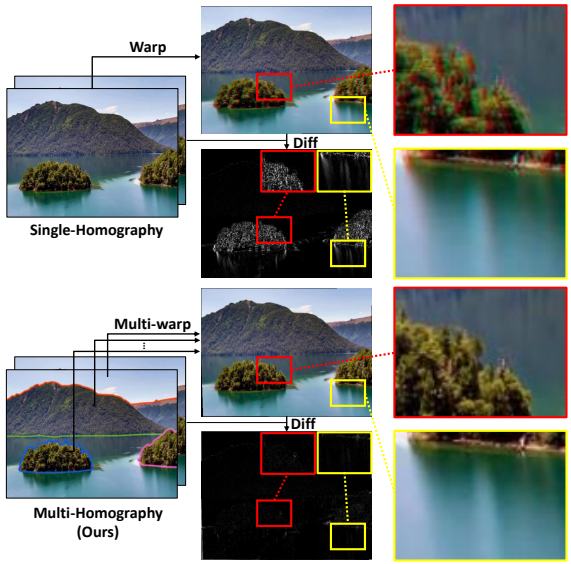

Figure 1: Comparison between single-H and multi-H. When multiple planes exist in a scene, single-H tends to focus on the dominant plane or find a balance between multiple planes, leading to misalignment in some regions. The overlay image is generated with R channel from the target image and G, B channels from the warped source image.

To deal with this, (Zhang et al. 2020) learns a mask to reject outlier regions and only select reliable regions for H estimation. Later, (Hong et al. 2022) proposes to guide the estimated H to focus on the dominant plane. Despite the efforts, these methods still obtain a global H, which is intrinsically a compromised result for multi-plane scenes.

Recently, (Liu et al. 2016) and (Nie et al. 2022) propose to estimate multiple Hs by dividing the image into mesh grids and computing a local H for each grid. (Liu et al. 2022b) and (Liu et al. 2022a) also propose to use mesh grid H estimation. By dividing the image into even mesh grids, there will be both cases where one grid contains multiple

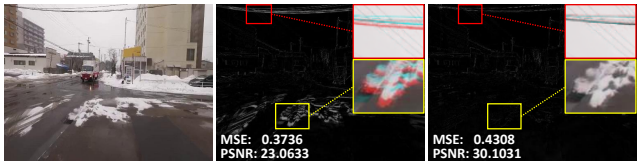

Figure 2: Comparison of different alignment effects evaluated using MSE(↓) and Mask-PSNR(↑). It can be noticed that in this case, compared to MSE, Mask-PSNR is more consistent with human observation.

planes and cases where one plane is divided into multiple grids, hence damaging the plane distribution, and the limitations of single-H models are also not fully addressed. Besides, calculated Hs for adjacent grids are likely not consistent when their dominant planes are different.

Another issue is that when evaluating H estimation performance, most existing works use Mean Squared Error (MSE), also referred to as Point Matching Error in (Zhang et al. 2020; Hong et al. 2022; Liu et al. 2022a). MSE calculates the deviations between the labeled matching point pairs within image pairs. However, as illustrated in Fig 2, a better MSE on sparse matching point pairs may not always guarantee better alignment of the entire image. In (Nie et al. 2022), the authors employ Peak Signal-to-Noise Ratio (PSNR), which calculates pixel-wise difference. Though, PSNR on the whole image can be affected by the plane-induced depth disparity and moving objects (people, vehicle, etc.) and cannot adequately represent the alignment quality.

To solve the aforementioned issues, we propose a novel framework, named **Mask-Homo**, for multi-H estimation and a new metric **Mask-PSNR** for more comprehensive evaluation of the alignment quality. In the proposed framework, given a pair of images, we obtain mask regions which correspond to pseudo planes within the images and carry out regional H estimation for the correlated mask pairs. The final warping output is obtained by fusing multiple warped images using estimated regional Hs. To summarize, our main contributions are as follows:

- A multi-H estimation framework, Mask-Homo, which solves the plane-induced depth disparity issue.
- A pseudo plane mask generation module, which obtains pseudo plane masks for regional H estimation, based on semantic information guidance.
- An auxiliary metric, Mask-PSNR, for more dense and visual consistent alignment quality evaluation.

Related Work

Image Segmentation Image segmentation methods can be broadly classified into three categories: instance segmentation (Li et al. 2017; Lee and Park 2020), semantic segmentation (Strudel et al. 2021; Hamilton et al. 2022), and panoptic segmentation (Li et al. 2019; Zhou et al. 2022). Semantic segmentation assigns a category label to each pixel to identify objects, instance segmentation focuses on identifying and segmenting individual instances of the objects, and panoptic segmentation combines the strengths of the previous two. Conventional segmentation aims to identify objects, while in our context, we aim for correlated image regions from the same plane within image pairs, which can be approximately induced by a homography.

Single-H Estimation Traditional approaches for H estimation typically involve detecting and matching feature points, rejecting outliers, and obtaining H with Direct Linear Transformation (Hartley and Zisserman 2003). With the advancement of deep learning, (DeTone, Malisiewicz, and Rabinovich 2016) introduces the first deep H estimation model in 2016, since when numerous methods have been proposed. Supervised methods (Le et al. 2020; Shao et al. 2021) use a synthetic dataset for training, that lacks realistic scene depth disparity, and generalize poorly on real images. In contrast, the Unsupervised method (Nguyen et al. 2018) uses real image pairs and develops an end-to-end algorithm by computing photometric loss. (Jiang et al. 2023) further proposes to generate a realistic dataset from unlabelled real-world image pairs. (Zhang et al. 2020) and (Le et al. 2020) propose to predict a mask to remove outliers and moving objects; while (Hong et al. 2022) proposes to guide the model to focus on the dominant plane by imposing a coplanarity constraint. Some SOTA image stitching methods (Nie et al. 2021, 2023) also explore different H estimation strategies. Although these methods have achieved good performance, they result in a global H, which is either a trade-off between multiple planes or focusing on the dominant plane, and still face the model inadequacy problem.

Multi-H Estimation To better handle the depth disparity challenge in multi-plane scenes, various approaches have been proposed. (Gao, Kim, and Brown 2011) proposes a dual H method which accounts for the distant plane and ground plane, separately. (Zaragoza et al. 2013) estimates a global projective warp while accommodating local deviations. (Lee and Sim 2020) partitions the image into super pixels and conducts warping based on a locally optimal H. However, these methods are feature-based and not robust in low texture scenes. For deep learning-based solutions, (Liu et al. 2016) introduces MeshFlow to predict a sparse motion field by dividing the image into mesh grids and computing a local H for each grid. (Liu et al. 2022b) and (Liu et al. 2022a) generalize the previous single-H method to local mesh grid H estimation. (Nie et al. 2022) also proposes to predict multi-grid H from global to local. Although these approaches are able to describe nonlinear motions better, the mesh grid separation of the image damages the plane distribution of the scene, and does not essentially handle the model inadequacy problem of single-H.

Optical Flow Optical flow (OF) (Sun et al. 2018; Teed and Deng 2020) is a different type of image alignment method from H. OF achieves heavy, pixel-level fine alignment with a high degree of freedom (DoF), while parametric H achieves light-weight, globally optimal alignment with a much lower DoF. This paper targets on H-based solutions. Multi-H offers a trade-off between the number of H and how much aligned is the image pair, and the goal is to obtain the best set of H to minimize the geometric errors.

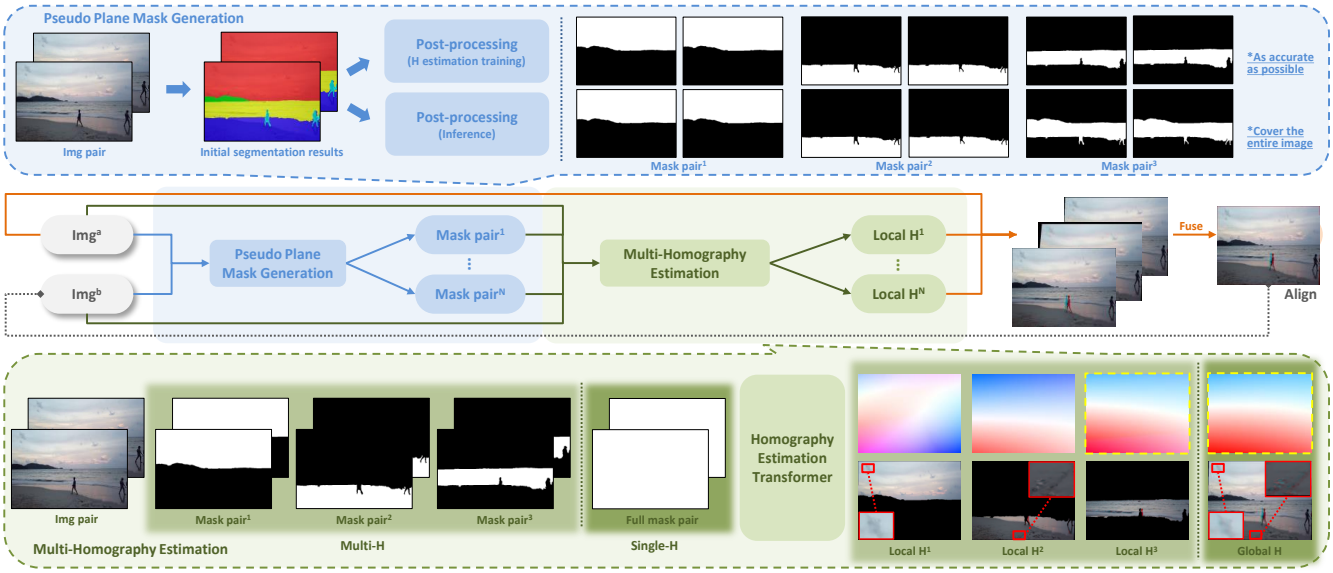

Figure 3: The overall pipeline of Mask-Homo. 1) Pseudo plane mask generation module obtains region correlations within the image pair. 2) Multi-homography estimation module predicts multiple local homographies for different regions. The multiple warped images are further fused to obtain the final result. It is interesting to find that the optical flow of the global H is very similar to one of the local Hs, as shown in the yellow dotted frame; this also proves that the global H estimation usually concentrates on one dominant plane instead of the whole image.

Method

Overview

The pipeline of the proposed Mask-Homo framework is illustrated in Fig 3. The framework has two main modules. The Pseudo Plane Mask Generation Module takes a pair of images I^a and I^b as input and outputs two sets of correlated pseudo plane mask pairs (Eq. 1).

$$(\mathbb{M}^a, \mathbb{M}^b) = \mathcal{S}(I^a, I^b) \tag{1}$$

The Multi-Homography Estimation Module takes the correlated mask pairs and image pair as input and outputs regional H estimation results for different regions (Eq. 2).

$$\mathbb{H} = \mathcal{H}(I^a, I^b; \mathbb{M}^a, \mathbb{M}^b) \tag{2}$$

Last, multiple warped images obtained with different regional Hs ($\mathbb{J}_{\mathbb{H}}^I = \Psi(\mathbb{H}, I)$) are fused to generate the final artifact-free output.

The notations are as follows: we use $\mathcal{S}$ and $\mathcal{H}$ to denote the two main modules, $I, M, H/\hat{H}, N$ to denote image, mask, homography for forward/backward warping and the number of regional homographies. Ψ is used to represent the warping operation and J_H is the warped image or feature by H. Blackboard bold font ($\mathbb{M}$ and $\mathbb{H}$) is used to represent sets.

Pseudo Plane Mask Generation

The goal of this module is to find region correlations between an image pair, where the two correlated regions can be approximately induced by a homography. The correlated regions include not only rigid planes such as ground or lake surface, but also planes in a more approximate sense, such as a range of buildings or mountains in the distance; we refer to them as pseudo planes. Intuitively, the masks for pseudo planes should be reasonably large, connected, and correlate to each other between the image pair, to enable robust and accurate local H estimation. We therefore form the fundamental geometric requirements for pseudo plane masks: with decent degree of connectivity and area.

Since connected pixels of same object category usually lie on same pseudo plane, we utilize semantic segmentation (Hamilton et al. 2022) to obtain initial segmentation results. However, they may be fragmented and not accurate in some regions, as shown in Image pair 1 of Fig 4(b). We conduct post-processing to acquire utilizable pseudo plane masks.

The required correlated masks for H estimation training and inference are slightly different. For H estimation training, we aim for mask pairs that are **as accurate as possible**. That is to say, we only focus on meaningful and credible regions to calculate a local H within an image pair. Specifically, we choose to trust the segmentation results with larger areas for different categories and rely on mask matching to reduce the influence of segmentation errors. Small or unmatched regions are not included for H estimation training. While for inference, we aim for mask pairs that **cover the entire image**. Specifically, for regions that are not credible enough to calculate a H, we assume that it is more likely to share the nearby local H.

The visual demonstration is shown in Fig 4. As shown in Image pair 2 of Fig 4(d)(e), for the yellow mask pair, when used for H estimation training, only the sea region mask pair is used; while for inference, nearby small masks including the mountain in the distance and persons are merged with it to form a larger mask pair that share the same local H.

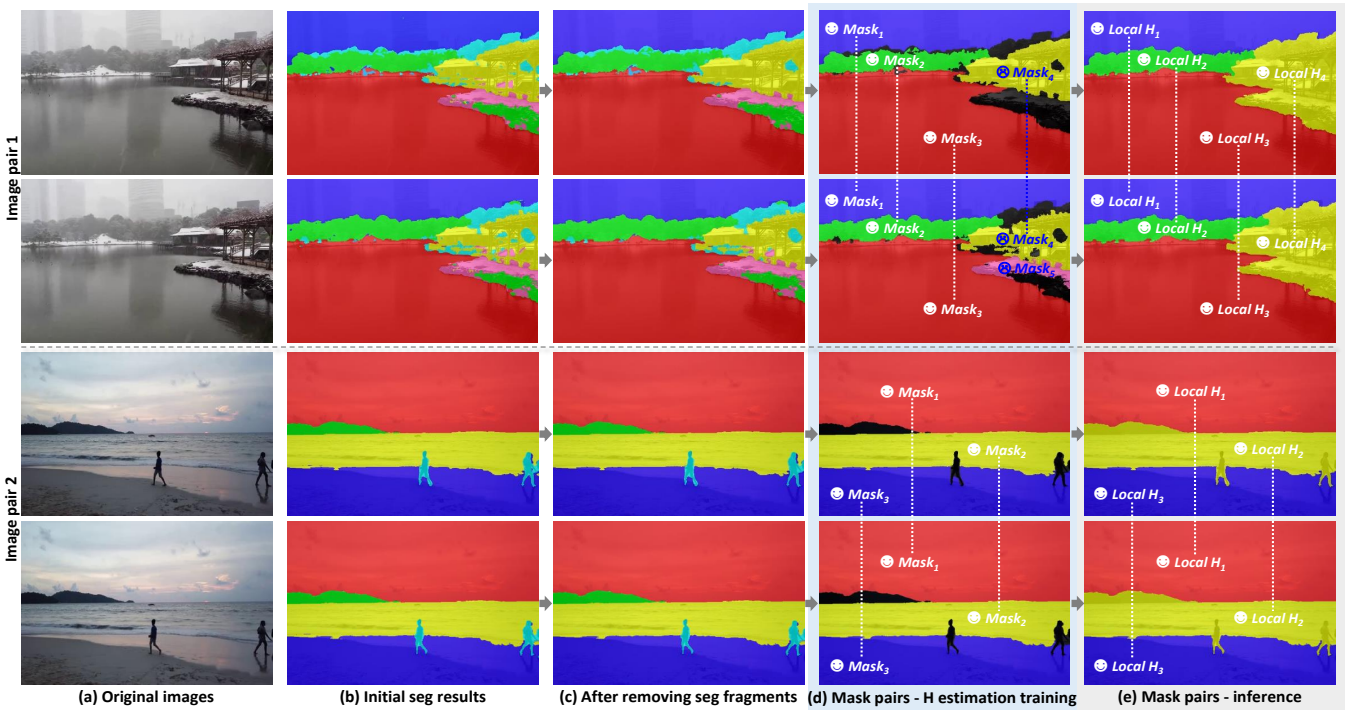

Figure 4: From left to right are original images, initial segmentation results, segmentation results after removing segmentation fragments, mask pairs for H estimation training and inference, respectively. For H estimation training, matched mask pairs share the same segmentation category, have close area and center point location; matched and unmatched mask pairs are represented with white happy face and blue sad face. Same marks are used to indicate whether local or global H is used for inference.

When there exists unmatched masks, they are combined and handled with global H, which is estimated from the entire image using full mask pair (bottom middle part in Fig 3).

The pseudo code of this module is depicted in Algorithm 1. We denote initial segmentation, mask matching operations as $\mathcal{K}$ and $\mathcal{P}$, size and class as S, C. S_{hole}, S_{seg}, D_{min}, S_{rto} and N are five hyper-parameters. Specific procedures of segmentation post-processing are presented below.

For H estimation training (1) Remove segmentation fragments that are smaller than S_{hole}. (2) Select masks that are larger than S_{seg}. (3) Remove moving object classes. (4) Mask matching. Two masks match only when the following conditions are true: (a) they belong to the same class; (b) center point location difference is minimal and lower than D_{min}; (c) size difference is lower than S_{rto}.

For inference (1) Remove segmentation fragments that are smaller than S_{hole}. (2) Select at most N masks that are larger than S_{seg}. (3) Merge unselected masks with selected ones. (4) Mask matching and image matching. Use local H and global H for matched and unmatched image pair, respectively. For partly matched image pair, local and global H are used together, for matched and unmatched mask pairs.

Multi-Homography Estimation

As can be seen in Fig 1, multiple Hs are required in order to align the sky, mountain, water and islands accurately and respectively. Multi-Homography Estimation Module conducts regional H estimation, based on previously acquired pseudo plane mask pairs. As H is estimated locally from a certain and mostly irregular region within the image pair, traditional 4-point parameterization (DeTone, Malisiewicz, and Rabinovich 2016), characterized with 4-corner offsets, is not applicable. We use 8 orthogonal flow bases parameterization(Ye et al. 2021) for regional H representation.

We follow the transformer network design by (Hong et al. 2022) for H estimation while incorporating the pseudo plane mask information. The query-key correlation of transformers establishes better local correspondence for H estimation, compared to CNN-based alternatives. The input images are first converted to feature maps from a *feature extractor*. Then, feature maps at multiple levels are extracted from a *multi-scale CNN encoder*. Last, the resulting feature pyramids are utilized for coarse-to-fine H estimation, using a *transformer* with cascaded self-attention encoder and class-attention decoder blocks.

To incorporate pseudo plane mask information, the mask is multiplied with image feature, which is extracted from the *feature extractor*, before being fed into the *multi-scale CNN encoder* for H estimation. For more details about the transformer network, please refer to (Hong et al. 2022).

As for the loss function, we also integrate the pseudo plane mask information into triplet loss (Schroff, Kalenichenko, and Philbin 2015) and feature identity loss (Ye et al. 2021). Given an image pair (I^a, I^b) and a correlated mask pair (M^a, M^b), which corresponds to a local homography H, masked triplet loss encourages the

Algorithm 1: Pseudo Plane Mask Generation $\mathcal{S}$

Input: Image pair: I^a, I^b
Output: Correlated masks: $\mathbb{M}^a, \mathbb{M}^b$

$\mathbb{M}_{init} \leftarrow \mathcal{K}(I^a, I^b)$ // (Fig.4(b))
for M *in* $\mathbb{M}_{init}$ **do**
  if $S_m < S_{hole}$ **then**
    Merge M to surrounding M'
end
Result: $\mathbb{M}_{filled}$ // (Fig.4(c))
for M *in* $\mathbb{M}_{filled}$ **do**
  case *Homography estimation training* **do**
    if $S_m > S_{seg}$ & $C_m \neq$ *moving obj.* **then**
      Add M to $\mathbb{M}_c$ // Candidate masks
    $(\mathbb{M}^a, \mathbb{M}^b) \leftarrow \mathcal{P}(\mathbb{M}_c)$ // Mask matching (Fig.4(d))
  end
  case *Inference* **do**
    if $S_m > S_{seg}$ & ($M \in$ *largest* N *masks)* **then**
      Add M to $\mathbb{M}_s$ // Selected masks
    else
      Add M to $\mathbb{M}_u$ // Unselected masks
    end
    for M *in* $\mathbb{M}_u$ **do**
      $\mathbb{M}'_s \leftarrow$ Merge M to closest M_s
    end
    $(\mathbb{M}^a, \mathbb{M}^b) \leftarrow \mathcal{P}(\mathbb{M}'_s)$ // Mask matching (Fig.4(e))
  end
end

masked region in I^a to approach the corresponding masked region in I^b, while the difference between two masked regions is maintained. Masked feature identity loss enforces the *feature extractor* ($\mathcal{G}$) to be warp-equivalent.

For the masked triplet loss when warping from I^a to I^b, the anchor is defined as $M^b \odot \mathcal{G}(I^b)$, the positive is defined as $M^b \odot J_H^{\mathcal{G}(I^a)}$ and the negative is defined as $M^a \odot \mathcal{G}(I^a)$. $\odot$ denotes element-wise multiplication; vice versa for backward warping from I^b to I^a.

We use L_{FI}, L_{Tri_f} and L_{Tri_b} to denote feature identity loss, forward and backward triplet loss, and $\ddot{L}$ to denote loss being calculated on masks. The definition for masked feature identity loss is shown in Eq. 3. The total loss function for H estimation training is summarized in Eq. 4.

$$\begin{aligned} \ddot{L_{fi}} = &||M^b \odot J_H^{\mathcal{G}(I^a)} - M^b \odot \mathcal{G}(J_H^{I^a})|| \\ &+ ||M^a \odot J_{\hat{H}}^{\mathcal{G}(I^b)} - M^a \odot \mathcal{G}(J_{\hat{H}}^{I^b})|| \end{aligned} \tag{3}$$

$$\ddot{L_H} = \ddot{L_f} + \ddot{L_b} + \ddot{L_{fi}} \tag{4}$$

With previous two modules, we have obtained multiple regional H transformations corresponding to different pseudo planes in the scene. Multiple warped images can be obtain accordingly, each of which aligns a correlated region precisely. In the inference stage, the generated pseudo plane mask pairs cover the entire image. We take advantage of this mask information to maintain consistent warping within individual mask regions, and conduct fusion to finally obtain an artifact-free and natural-looking result.

Experiments

Dataset Our method is evaluated on a natural image dataset (Zhang et al. 2020; Liu et al. 2022a) with 75.8k training pairs and 4.2k testing pairs. The scenes in the dataset are roughly categorized into five types: REgular (RE), Low Texture (LT), Low Light (LL), Small Foreground (SF) and Large Foreground (LF), where the last four types are more challenging. For each test pair of images, 8-10 labeled matching point pairs are provided. Six of them are located on the dominant plane and can be used for global H evaluation, while the rest 2-4 point pairs are from other planes and can be further used for local H estimation.

Evaluation Metrics As aforementioned, apart from the conventional MSE, we further utilize Mask-PSNR for more comprehensive evaluation. In (Nie et al. 2022), PSNR is calculated on the overlapping regions after warping of the entire image. However, the existence of depth disparity and moving objects affects its accuracy. Thus, we propose Mask-PSNR, which calculates PSNR on the correlated mask regions. Mask-PSNR avoids the influence of depth disparity by following the region correlation hypothesis, and the effect of moving objects by segmentation post-processing.

Implementation Details For training, we randomly crop 384×512 patches near the center of original images to avoid out-of-bound coordinates after warping. Other parameters for H estimation transformer are same to (Hong et al. 2022). Adam optimizer (P. Kingma and Ba 2015) is employed. For H estimation training, the learning rate is 1×10^{-4}, which decays by a factor of 0.8 after every epoch, batch size is 8 and it takes 10 epochs to train.

For segmentation post-processing, there are five hyper-parameters involved: S_{hole}, S_{seg}, D_{min}, S_{rto} and N. The first four parameters determine the shape of generated segmentation masks. We empirically find that D_{min} affects the diversity of generated segmentation masks much more significantly than others. Thus in our experiments, we fix $S_{hole} = 500$, $S_{seg} = 10,000$, $S_{rto} = 15\%$, while D_{min} is varied to investigate the influence of segmentation post-processing on the performance. The last parameter N decides the maximum number of pseudo plane masks within each image pair in inference, which is empirically set to 4.

Comparison with Existing Methods

To qualitatively and quantitatively evaluate the performance of the proposed method, we report comparisons with 5 single-H methods: **Supervised** (DeTone, Malisiewicz, and Rabinovich 2016), **Unsupervised** (Nguyen et al. 2018), **CA-Unsupervised** (Zhang et al. 2020), **BasesHomo** (Ye et al. 2021), **HomoGAN** (Hong et al. 2022); 3 multi-H[1] methods: **APAP** (Zaragoza et al. 2013), **MeshFlow** (Liu et al. 2016), **MeshBasesHomo** (Liu et al. 2022a); 2 dense optical flow methods: **PWCNet** (Sun et al. 2018), **RAFT** (Teed and Deng 2020). Due to space limitations, more detailed results can be found in the supplementary material.

[1]MeshBasesHomo (Liu et al. 2022a) is the latest and SOTA multi-H work. However the code for the mesh grid H estimation part has not been released yet, thus we cannot test it with Mask-PSNR or conduct qualitative comparisons.

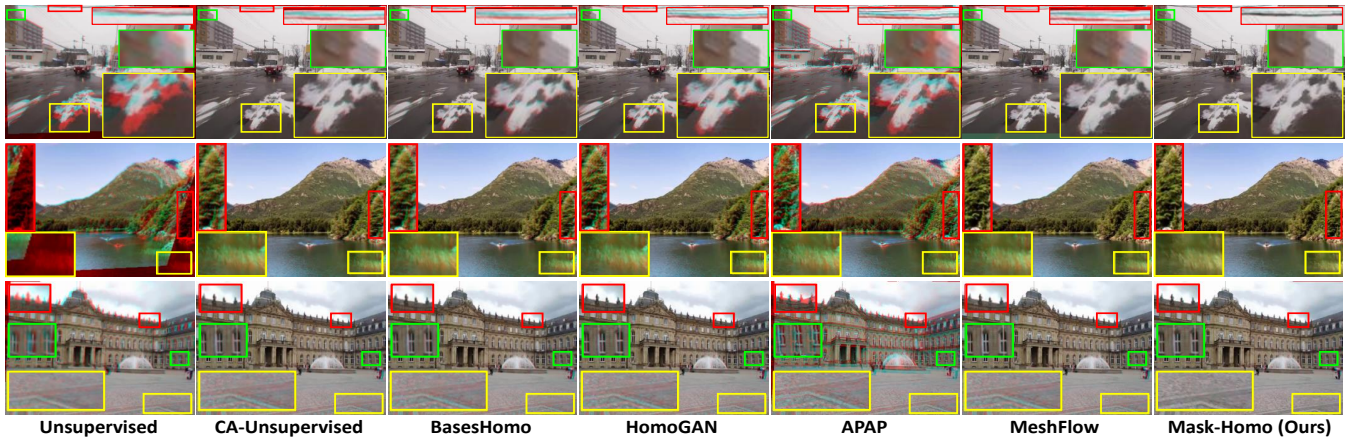

Figure 5: Qualitative results of our method and six other deep learning-based methods. First four are single-H methods; next two are multi-H methods. Error-prone regions are highlighted with red, yellow and green boxes. Best viewed with zooming in.

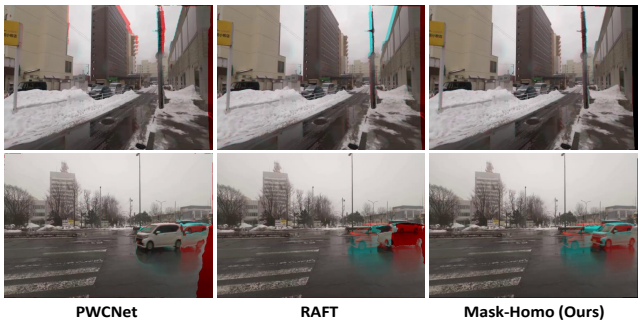

Figure 6: Qualitative results with two OF methods.

Qualitative comparison In Fig 5, we present qualitative results of our method together with six recent works. First four are single-H methods; next two are multi-H methods.

The first example is challenging as it contains a range of buildings, distant power lines and snow covered foreground pavement, resulting in diverse planes and depth disparities. The second and third examples also include multiple planes, of which the dominant plane locates at the mountain and building, respectively. As highlighted in colored boxes, existing methods cannot align these images as well as ours.

The Unsupervised method (Nguyen et al. 2018) predicts a single H based on the entire image, leading to a compromised result on multiple planes. CA-Unsupervised (Zhang et al. 2020) and BasesHomo (Ye et al. 2021) select reliable regions when estimating the H, while HomoGAN (Hong et al. 2022) focuses on the dominant plane. These methods perform well on aligning regions of concern, however the rest regions of the image are ignored, leading to low performance on whole image evaluation. For instance, as the yellow boxes in three examples are not located in dominant regions, none of previous methods can align them as well as ours. APAP and MeshFlow (Zaragoza et al. 2013; Liu et al. 2016) learn multiple Hs based on image mesh grids. However, Hs estimated from local mesh grids are not as accurate as from the proposed pseudo planes. Our method estimates multiple Hs following the plane distribution, and is able to align different regions simultaneously.

In Fig 6, we provide warping results using our method and OF. As can been seen, OF sometimes damages the consistency of the image content, or fails when moving object passes quickly. Our network learns a special OF constrained by the 8 H bases. It is embedded within a 8-D subspace, which is significantly smaller than $2HW$-D space of a general OF. In both cases, our method is able to align different regions, with a much lower DoF.

Quantitative comparison We report quantitative comparisons with 5 single-H methods, 3 multi-H methods, and 2 dense optical flow methods, using MSE and Mask-PSNR.

As can been seen from Table 1, the MSE in the upper half (Row 3-10) measures the error between 6 pairs of matching points on the dominant plane and is used for single-H evaluation. While, the lower half (Row 11-18) measures the error between all pairs of matching points, some of which are outside the dominant plane, therefore more suitable for multi-H evaluation. The Mask-PSNR reflects the similarity between the correlated region pair after warping.

For MSE, when compared with single-H methods (Row 4-9), our method achieves better performance than SOTA methods in most cases. Our method outperforms the baseline HomoGAN* (Hong et al. 2022) ($p.t.$) by 16% (0.49→0.41) and even the best HomoGAN* ($f.t.$) as well (0.42→0.41). However, for the LT scene, we are having slightly worse MSE. This may be because that our method estimates H from local regions, which is affected by low texture regions in these scenes. When compared with multi-H methods (Row 12-15), the SOTA method is MeshBasesHomo (Liu et al. 2022a). Our method outperforms it by 13% (0.79→0.69) when using similar amount of H (our method uses at most 4 H). Even for MeshBasesHomo with 8×8 mesh, i.e. 64 H, which is much larger than ours, our method still surpasses it in RE and LT cases. For the LF scene, our method does not perform as well, and we think it is related to that most LF scenes contain moving objects (cars, etc.) that occur as large foreground, and in our H estimation

	MSE (↓)						Mask-PSNR (↑)					
	RE	LT	LL	SF	LF	Avg	RE	LT	LL	SF	LF	Avg
$\mathcal{I}_{3\times3}$[2] - *dominant plane*	7.75	7.65	7.21	7.53	3.39	6.70	29.41	39.10	35.72	34.60	36.84	35.13
Supervised	1.51	4.48	2.76	2.62	3.00	2.87	26.07	31.41	33.20	27.70	27.52	29.18
Unsupervised	0.79	2.45	1.48	1.11	1.10	1.39	Inf[4]	30.73	32.86	26.97	Inf	Inf
CA-Unsupervised	0.73	1.01	1.03	0.92	0.70	0.88	37.54	38.06	42.08	35.35	35.14	37.64
BasesHomo	0.29	0.54	0.65	0.61	0.41	0.50	37.57	38.02	42.27	35.67	35.28	37.76
HomoGAN* (*p.t.*)[3]	0.28	0.49	0.61	0.62	0.45	0.49	40.31	42.96	43.28	38.75	39.91	41.04
HomoGAN* (*f.t.*)	0.26	**0.40**	0.60	0.49	0.32	0.42	40.27	42.79	43.25	38.51	39.95	40.95
Mask-Homo (Ours) - *dominant plane*	**0.22**	0.61	**0.54**	**0.39**	**0.31**	**0.41**	**40.47**	**43.20**	**43.32**	**40.76**	**41.87**	**41.93**
$\mathcal{I}_{3\times3}$ - *all points*	7.81	7.87	7.49	8.34	4.14	7.13	29.41	39.10	35.72	34.60	36.84	35.13
APAP	1.59	2.72	1.75	1.70	2.10	1.97	28.38	32.31	35.11	28.06	29.95	30.76
MeshFlow	0.46	1.04	1.06	1.09	1.36	1.00	37.50	37.53	41.72	35.14	37.35	37.85
MeshBasesHomo (2×2 mesh)	0.39	1.01	0.85	0.72	0.99	0.79	-[4]	-	-	-	-	-
MeshBasesHomo (8×8 mesh)	0.32	0.91	**0.67**	**0.48**	**0.74**	**0.62**	-	-	-	-	-	-
PWCNet	*0.42*	*1.51*	*0.82*	*1.03*	*0.99*	*0.95*	*32.54*	*31.98*	*41.60*	*34.81*	*33.70*	*34.93*
RAFT	*0.32*	*0.99*	*0.74*	*0.49*	*0.88*	*0.68*	*37.29*	*39.79*	*42.50*	*39.02*	*40.37*	*39.79*
Mask-Homo (Ours) - *all points*	**0.27**	**0.86**	0.73	0.55	1.05	0.69	**40.47**	**43.20**	**43.32**	**40.76**	**41.87**	**41.93**

Table 1: MSE and Mask-PSNR comparison results of our method with both traditional and deep learning-based, single-H and multi-H methods. In the upper half, MSE reports the mean squared error between six pairs of matching points that are located on the dominant plane and is used for single-H evaluation. In the lower half, MSE is between all pairs of matching points, some of which are located in areas outside the dominant plane, and more suitable for multi-H evaluation. The best and second best results are highlighted using boldface and underlining, respectively. Moreover, the results of two OF methods are added as reference, which are highlighted using italics.

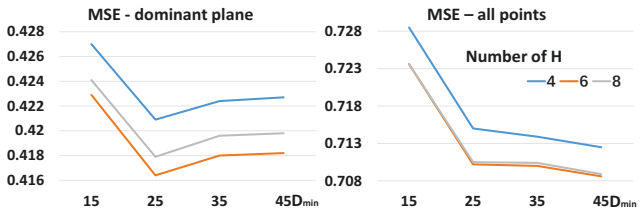

Figure 7: Investigation of the influence of pseudo plane mask generation on the homography estimation performance with two most influential parameters: D_{min} and N.

training, moving objects are removed. When compared with OF methods (Row 16-17), our method surpasses PWCNet and achieves comparable result to RAFT, which further verify the effectiveness of the proposed framework.

For Mask-PSNR, our method notably outperforms all existing methods (0.89dB, compared to SOTA), especially for SF and LF scenes (2.01dB and 1.92dB). This also agrees with our intuition that global H cannot align the images well when notable foreground or multiple planes exist.

Investigation of Pseudo Plane Mask Generation

In this subsection, we investigate the influence of pseudo plane mask generation on the performance. As introduced earlier, the pseudo plane mask generation is mainly affected by 2 parameters: D_{min} and N. D_{min} is maximal center point location difference. A smaller value guarantees more accurate mask matching, while a larger value allows less strict matching with larger diversity. N is maximum number of pseudo plane masks within image pairs. In experiments, we find that N being set to 4 is able to cover most cases.

Fig 7 shows the performance with different settings of D_{min} and N. MSE results on six dominant plane point pairs and all point pairs are displayed, respectively. The figures present how performance changes when D_{min} is varied from 15 to 45 with an interval of 10, and when N is varied from 4 to 8 with an interval of 2. With a larger value of D_{min}, which indicates more loose mask matching and more paired of local regions being found, the performance first increases and then stays stable. This shows the effectiveness of the multi-H framework. With an increasing number of H being used, a performance gain can also be observed. Despite the performance change, the results with different parameter settings are all high and stable, indicating the robustness of the proposed pseudo plane mask generation.

Conclusion

A major challenge in multi-H estimation is how to obtain correlated regions that follow the scene plane distribution. In this paper, we explore a novel framework for this task. By incorporating local pseudo plane mask information, which is obtained in an unsupervised manner, we achieve better local alignment; and further obtain a globally fused natural-looking result. Experiments prove that the proposed method can qualitatively and quantitatively achieve better alignment when compared with SOTA.

For future work, (1) jointly considering depth and semantics; (2) an end-to-end deep learning framework that learns correlated regions and H simultaneously will be our target.

[2] $\mathcal{I}_{3\times3}$ refers to identity transformation.

[3] * denotes reproduced results using officially released pre-train(p.t.) and fine-tune(f.t.) models. HomoGAN*(p.t.) is the baseline model we utilize.

[4] Inf means existence of mask regions having no overlapping.

References

DeTone, D.; Malisiewicz, T.; and Rabinovich, A. 2016. Deep Image Homography Estimation. *arXiv preprint arXiv: 1606.03798*.

Gao, J.; Kim, S. J.; and Brown, M. S. 2011. Constructing image panoramas using dual-homography warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 49–56.

Hamilton, M.; Zhang, Z.; Hariharan, B.; Snavely, N.; and Freeman, W. T. 2022. Unsupervised Semantic Segmentation by Distilling Feature Correspondences. In *Proc. ICLR*.

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

Hong, M.; Lu, Y.; Ye, N.; Lin, C.; Zhao, Q.; and Liu, S. 2022. Unsupervised Homography Estimation with Coplanarity-Aware GAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17642–17651.

Jiang, H.; Li, H.; Han, S.; Fan, H.; Zeng, B.; and Liu, S. 2023. Supervised Homography Learning with Realistic Dataset Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9806–9815.

Le, H.; Liu, F.; Zhang, S.; and Agarwala, A. 2020. Deep Homography Estimation for Dynamic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7649–7658.

Lee, K.-Y.; and Sim, J.-Y. 2020. Warping Residual Based Image Stitching for Large Parallax. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8195–8203.

Lee, Y.; and Park, J. 2020. Centermask: Real-time anchor-free instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13906–13915.

Li, Y.; Chen, X.; Zhu, Z.; Xie, L.; Huang, G.; Du, D.; and Wang, X. 2019. Attention-guided unified network for panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7026–7035.

Li, Y.; Qi, H.; Dai, J.; Ji, X.; and Wei, Y. 2017. Fully convolutional instance-aware semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2359–2367.

Liu, S.; Lu, Y.; Jiang, H.; Ye, N.; Wang, C.; and Zeng, B. 2022a. Unsupervised Global and Local Homography Estimation With Motion Basis Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, S.; Tan, P.; Yuan, L.; Sun, J.; and Zeng, B. 2016. Meshflow: Minimum latency online video stabilization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, 800–815. Springer.

Liu, S.; Ye, N.; Wang, C.; Zhang, J.; Jia, L.; Luo, K.; Wang, J.; and Sun, J. 2022b. Content-Aware Unsupervised Deep Homography Estimation and its Extensions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 2849–2863.

Nguyen, T.; Chen, S. W.; Shivakumar, S. S.; Taylor, C. J.; and Kumar, V. 2018. Unsupervised Deep Homography: A Fast and Robust Homography Estimation Model. *IEEE Robotics and Automation Letters*, 3(3): 2346–2353.

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

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

Nie, L.; Lin, C.; Liao, K.; Liu, S.; and Zhao, Y. 2023. Parallax-Tolerant Unsupervised Deep Image Stitching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7399–7408.

P. Kingma, D.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *Proc. ICLR*.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Shao, R.; Wu, G.; Zhou, Y.; Fu, Y.; Fang, L.; and Liu, Y. 2021. LocalTrans: A Multiscale Local Transformer Network for Cross-Resolution Homography Estimation. In *IEEE/CVF International Conference on Computer Vision*, 14870–14879.

Strudel, R.; Garcia, R.; Laptev, I.; and Schmid, C. 2021. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7262–7272.

Sun, D.; Yang, X.; Liu, M.-Y.; and Kautz, J. 2018. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8934–8943.

Teed, Z.; and Deng, J. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, 402–419. Springer.

Ye, N.; Wang, C.; Fan, H.; and Liu, S. 2021. Motion Basis Learning for Unsupervised Deep Homography Estimation with Subspace Projection. In *IEEE/CVF International Conference on Computer Vision*, 13097–13105.

Zaragoza, J.; Chin, T.-J.; Brown, M. S.; and Suter, D. 2013. As-Projective-As-Possible Image Stitching with Moving DLT. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2339–2346.

Zhang, J.; Wang, C.; Liu, S.; Jia, L.; Ye, N.; Wang, J.; Zhou, J.; and Sun, J. 2020. Content-Aware Unsupervised Deep Homography Estimation. In *European Conference on Computer Vision*, 653–669. ISBN 978-3-030-58452-8.

Zhou, Y.; Zhang, H.; Lee, H.; Sun, S.; Li, P.; Zhu, Y.; Yoo, B.; Qi, X.; and Han, J.-J. 2022. Slot-vps: Object-centric representation learning for video panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3093–3103.