Towards Generalized Multimodal Homography Estimation

Jinkun You, Jiaxin Cheng, Jie Zhang, Yicong Zhou*
Department of Computer and Information Science, University of Macau
Macau, China
youjinkun09@gmail.com, yc47434@um.edu.mo, jiezh1997@gmail.com, yicongzhou@um.edu.mo

Abstract

Supervised and unsupervised homography estimation methods depend on image pairs tailored to specific modalities to achieve high accuracy. However, their performance deteriorates substantially when applied to unseen modalities. To address this issue, we propose a training data synthesis method that generates unaligned image pairs with ground-truth offsets from a single input image. Our approach renders the image pairs with diverse textures and colors while preserving their structural information. These synthetic data empower the trained model to achieve greater robustness and improved generalization across various domains. Additionally, we design a network to fully leverage cross-scale information and decouple color information from feature representations, thus improving estimation accuracy. Extensive experiments show that our training data synthesis method improves generalization performance. The results also confirm the effectiveness of the proposed network.

1. Introduction

Homography estimation identifies a matrix that characterizes the projective transformation between two images of the same scene captured from different viewpoints. This transformation allows one image to be warped using the estimated matrix, enabling spatial alignment with the other image. Achieving this alignment is crucial for various applications, such as image stitching [4, 20, 25], image fusion [33, 46, 47], and guided super-resolution [32, 37, 39].

Traditional estimation methods depend on hand-crafted features to match points between image pairs for homography matrix calculation [16, 22, 28]. However, these approaches yield insufficient features in low-texture scenarios. To overcome this limitation, deep learning techniques have been introduced, leveraging their remarkable capabilities for feature extraction and representation across various domains [11, 26, 38]. The pioneering deep homography estimation method, inspired by supervised learning, extracts features from image pairs to predict the positional offsets of four corner points [12]. These offsets can then be used to compute the homography matrix through the direct linear transform algorithm [13]. During training, the model's weights are adjusted by minimizing the difference between the predicted offsets and the ground truth. To further reduce estimation errors, recent supervised approaches have adopted the inverse compositional Lucas-Kanade framework to refine the predicted offsets iteratively [2, 5, 45]. These methods estimate offsets multiple times and progressively aggregate the results. Nevertheless, supervised methods require well-aligned image pairs or unaligned pairs with ground-truth offsets for training. The collection of aligned pairs poses a significant challenge, particularly for multimodal images captured by different sensors [19]. Additionally, obtaining ground-truth data in real-world scenarios is often difficult [24]. These challenges substantially impede the training of supervised methods.

Unsupervised methods have gained significant attention recently because they do not require aligned images or ground truth for training. Early approaches focused on optimizing model parameters by maximizing visual similarity between image pairs [23, 36, 42]. To achieve this, offsets are predicted to align the images prior to calculating their similarity. However, these methods struggle with accuracy on multimodal image pairs due to substantial differences in appearance. Some strategies employ image translation techniques to standardize modalities while simultaneously learning homography estimation [1, 31, 34]. Nonetheless, these methods are limited to handling only small deformations, resulting in low accuracy. Self-supervised learning techniques have been introduced to address this issue. These methods generate pseudo-unaligned image pairs along with the corresponding ground truth, thereby enhancing homography estimation capabilities [41, 43]. The modality gap can be narrowed by enforcing consistency in image appearance or feature representation [30, 41, 43]. As a result, the estimation results are further improved.

Supervised and unsupervised homography estimation

*Corresponding author.

methods have demonstrated promising accuracy. As illustrated in Fig. 1, these models perform admirably when trained and tested on the same dataset. The training data helps adapt the models to the specific modalities involved. However, there are several limitations. First, the trained models exhibit limited generalization capability to other modalities. The estimation accuracy declines when image pairs demonstrate significant appearance differences across modalities. To improve performance, existing approaches require the collection of numerous image pairs from target modalities to learn modality-specific information. This need escalates both time expenditure and labor costs. Second, current methods tend to use features from different scales in isolation. While they effectively harness intra-scale information, they neglect complementary cross-scale information that is beneficial for establishing correspondences between the image pair [40]. Third, the integration of color information within the features can degrade the processing capabilities for multimodal images [27].

To address the above challenges, we introduce a training data synthesis method along with a homography estimation network. Our contributions are as follows:

- We propose a training data synthesis method that enables zero-shot multimodal homography estimation. By generating synthetic data with various textures and colors, our approach enhances models' ability to generalize across different modalities. This synthesis method can also be applied to existing datasets, leading to improved generalization performance.
- We design a homography estimation network to achieve higher accuracy. It integrates cross-scale information in both top-to-bottom and bottom-to-top directions. Besides, the network decouples color information from feature representations, further enhancing the estimation.
- We conduct comprehensive experiments to demonstrate the effectiveness of both the proposed training data synthesis method and the homography estimation network.

2. Related Work

Supervised Homography Estimation. DeTone *et al.* [12] found that parameterizing homography estimation as four-point offsets is more effective. They developed a VGG-style network to regress these offsets based on concatenated image pairs. The model is trained by minimizing the difference between the ground truth and the predicted offsets. Existing supervised methods adopt this training framework. However, regressing offsets just once limits accuracy. To address this, some approaches [14, 18, 48] cascade multiple networks for warping and processing input images multiple times. This strategy progressively refines the estimation result and reduces estimation errors, but it also introduces additional parameters. In contrast, the inverse compositional Lucas-Kanade (IC-LK) algorithm maintains model compactness while allowing iterative refinement of the estimation results. Chang *et al.* [7] introduced a differentiable yet untrainable IC-LK layer, enabling effective model training. Building on this, Le *et al.* [45] enhanced performance for multimodal image pairs through brightness-consistent feature extraction. Cao *et al.* [5] proposed a trainable iterative framework to fully leverage the data-driven capabilities of deep learning. They subsequently designed an attention mechanism to capture correspondences across various ranges [6]. Zhu *et al.* [49] integrated the iterative strategy into multiscale correlation searching, resulting in improved estimation accuracy and efficiency.

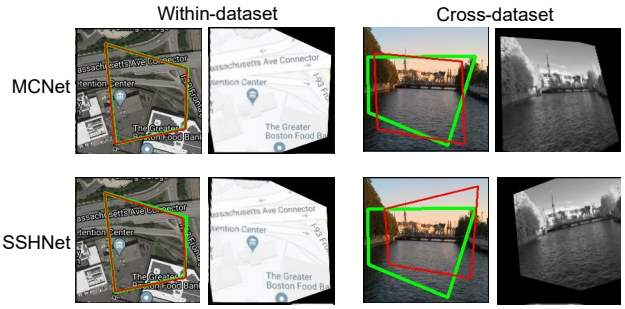

Figure 1. Estimation results. The rows present the results of the supervised MCNet [49] and the unsupervised SSHNet [41]. The first two columns represent within-dataset results, while the remaining columns depict cross-dataset performance. Greater similarity between the green and red quadrilaterals indicates higher accuracy.

Unsupervised Homography Estimation. Ground-truth offsets are often unavailable in real-world scenarios. This issue prompts increased interest in unsupervised methods since they do not require ground truth data for model training. Nguyen *et al.* [23] introduced the first unsupervised method. This method featured a differentiable warping layer and utilized photometric loss to minimize the visual discrepancies between image pairs. To enhance estimation accuracy, Zhang *et al.* [42] implemented attention maps to remove outliers and maximized similarity for aligned feature maps. Ye *et al.* [36] developed a warp-equivariant feature extractor to reduce the rank of the homography flow. However, these methods are primarily constrained to unimodal images. While some methods have utilized modality transfer to work with multimodal image pairs, they have achieved limited accuracy. To address this challenge, Zhang *et al.* [43] proposed a self-supervised framework that generates unaligned image pairs and corresponding ground truth from each training image. These generated data significantly enhance the unimodal estimation capabilities. The features [30, 43] or modalities [41] of the image pair are aligned to allow for multimodal homography estimation.

Style Transfer. A style transfer network merges the style of a given template image with the content of another im-

age. Gatys *et al.* [15] introduced the first convolutional neural network capable of extracting style features and integrating them with content features. Chen *et al.* [8] designed an internal-external network to tackle issues of color imbalance and pattern repetition. Cheng *et al.* [9] developed a loss function to reduce model bias toward specific styles. Zhang *et al.* [44] enhanced style capture by leveraging text prompts within a pretrained diffusion model. Chung *et al.* [10] transformed self-attention into cross-attention, eliminating the need to train diffusion models.

3. Proposed Method

3.1. Overall Framework

Fig. 2 illustrates the proposed training data synthesis method and the homography estimation network. The synthetic data are utilized for model training to achieve zero-shot estimation and enhance generalization performance. The model training is formulated as:

$$\theta^* = \max_{\theta} \mathrm{P}(\mathrm{Net}(I_{src}, I_{tar}, \theta), O_{gt}), \tag{1}$$

where $I_{src} \in \mathbb{R}^{3\times S\times S}$, $I_{tar} \in \mathbb{R}^{3\times S\times S}$, and $O_{gt} \in \mathbb{R}^{4\times 2}$ denote the synthetic source image, target image, and groud-truth offsets, respectively; $\mathrm{Net}(\cdot,\cdot)$ denotes the estimation model and θ represents its parameters; $\mathrm{P}(\cdot,\cdot)$ evaluates the estimation accuracy. The final objective is given by

$$\max \mathrm{P}(\mathrm{Net}(I'_{src}, I'_{tar}, \theta^*), O'_{gt}), \tag{2}$$

where I'_{src} and I'_{tar} are the image pairs with unseen modalities; O'_{gt} is the ground-truth offsets. The proposed network integrates cross-scale information and decouples color information for higher estimation accuracy. It also employs the iterative strategy to refine the estimation results.

3.2. Training Data Synthesis

As observed in Fig. 1, textures and colors vary across different modalities. Therefore, the homography estimation model must be robust to changes in textures and colors. Style transfer networks can effectively blend texture and color information from one image into another, inspiring us to use this approach to render the same image in various styles. These rendered images are then utilized for training data synthesis, providing a diverse array of textures and colors while preserving the structural information.

The synthetic training data consist of the source image $I_{src} \in \mathbb{R}^{3\times S\times S}$, the target image $I_{tar} \in \mathbb{R}^{3\times S\times S}$, and the ground-truth offsets $O_{gt} \in \mathbb{R}^{4\times 2}$. Initially, a content image $I_c \in \mathbb{R}^{3\times H\times W}$ is randomly sampled from the content dataset $\mathcal{X} = \{I_c^1, I_c^2, ..., I_c^N\}$. A patch is then obtained by cropping I_c

$$I_{patch} = \mathrm{Crop}(I_c, x, y, S_m + S), \tag{3}$$

where S_m is the margin size; $I_{patch} \in \mathbb{R}^{3\times(S_m+S)\times(S_m+S)}$ represents the cropped patch with its top-left corner positioned at (x, y) in I_c; $\mathrm{Crop}(\cdot)$ is the cropping function. Next, aligned source and target images are generated by rendering I_{patch} in different styles. Two template images, I_t^i and I_t^j, are randomly selected from the template dataset $\mathcal{Y} = \{I_t^1, I_t^2, ..., I_t^M\}$. The pairs (I_{patch}, I_t^i) and (I_{patch}, I_t^j) are fed into the style network as follows:

$$I_{src} = \alpha_i \cdot I_{patch} + (1-\alpha_i) \cdot \mathrm{Net}_s(I_{patch}, I_t^i), \tag{4}$$

$$I_{tar} = \alpha_j \cdot I_{patch} + (1-\alpha_j) \cdot \mathrm{Net}_s(I_{patch}, I_t^j), \tag{5}$$

where $\mathrm{Net}_s(\cdot)$ denotes the style transfer network; α_i and α_j are content weights uniformly distributed over the interval $[0, 1]$. A larger value indicates greater similarity to I_{patch}. Since the style transfer network does not control the smoothness of textures, we apply image smoothing [35] to I_{src} and I_{tar} by

$$I_{src}, I_{tar} = \mathrm{Smooth}(I_{src}, \beta_i), \mathrm{Smooth}(I_{tar}, \beta_j), \tag{6}$$

where β_i and β_j are smoothing weights uniformly distributed within $[0, \beta]$. A higher value results in greater smoothness. Subsequently, the ground-truth offsets O_{gt} are generated to warp I_{src}

$$I_{src} = \mathrm{Warp}(I_{src}, O_{gt}), \tag{7}$$

where $\mathrm{Warp}(\cdot)$ is the homography transformation function. Note that each element in O_{gt} belongs to the integer set $\{-p, -p+1, ..., p\}$, with p representing the maximum perturbation. Finally, center patches of size $S \times S$ are extracted from I_{src} and I_{tar} to create unaligned image pairs. Homography estimation models are trained on these unaligned image pairs along with the ground-truth offsets in a supervised manner. The content and template datasets consist solely of unimodal color images.

3.3. Cross-Scale and Color-Invariant Network

Existing homography models extract multiscale features to iteratively refine estimation results. However, these models focus solely on intra-scale information while neglecting cross-scale information, which limits accuracy. Additionally, they integrate color information from the image pairs into the extracted features, thus degrading generalization performance. To address these challenges, we propose the Cross-Scale and Color-Invariant Network (CCNet). CCNet effectively fuses features from different scales and decouples color information from the fused features. It utilizes the iterative strategy to reduce estimation errors.

Overall Structure. Fig. 2(b) shows the structure of CCNet. For $\forall i \in \{src, tar\}$, the multiscale features of I_i are extracted by

$$F_i^1, F_i^2, F_i^3 = \mathrm{Extractor}(I_i), \tag{8}$$

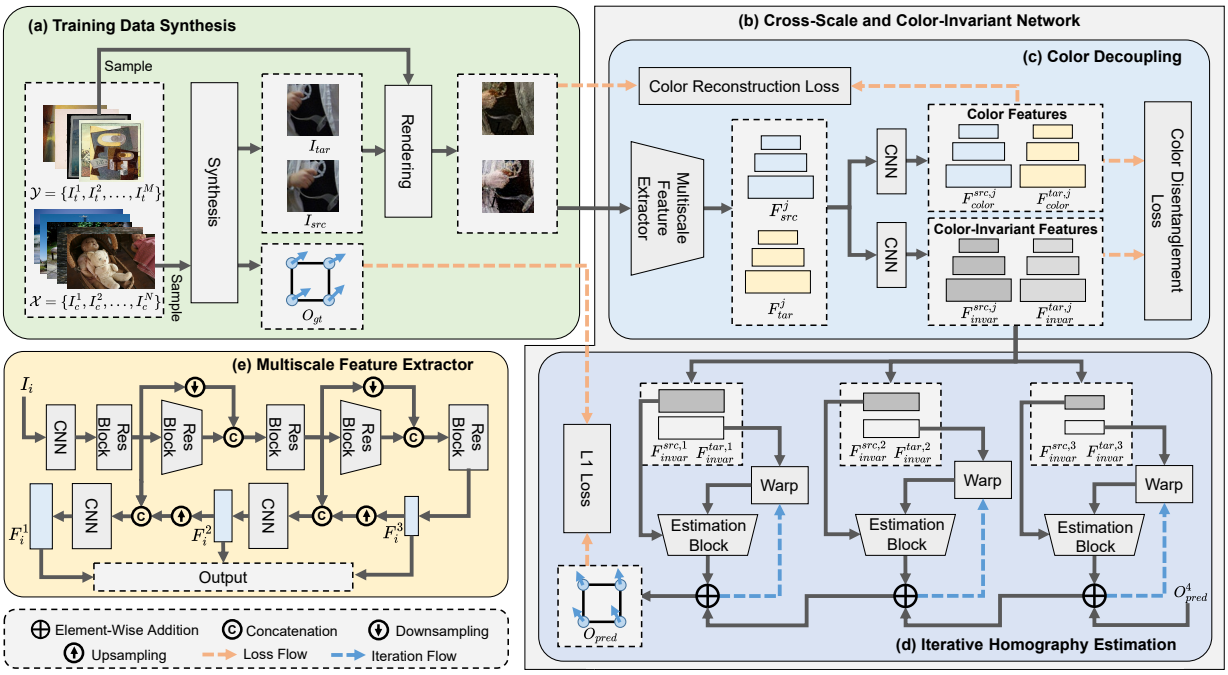

Figure 2. Illustration of the training data synthesis and the homography estimation network. (a) Training Data Synthesis can be applied to a public RGB dataset for zero-shot homography estimation or integrated with existing datasets to enhance generalization. (b) Cross-Scale and Color-Invariant Network (CCNet) integrates cross-scale information into the extracted features while decoupling color information from the feature representations. (c) Color Decoupling. (d) Iterative Homography Estimation. (e) Multiscale Feature Extractor.

where F_i^1, F_i^2, and F_i^3 have spatial sizes of $S \times S$, $S/2 \times S/2$, $S/4 \times S/4$, respectively; $\text{Extractor}(\cdot)$ is a multiscale feature extraction network fusing cross-scale information. These extracted features are then passed through convolutional layers designed to eliminate color information. The resulting features are color-invariant and fed into the iterative homography estimation module for predicting offsets.

Multiscale Feature Extractor. Fig. 2(e) illustrates the architecture of the extractor. For $\forall i \in \{src, tar\}$, the input image I_i is processed through a convolutional layer and a residual block to extract the shallow feature $F^1 \in \mathbb{R}^{C \times S \times S}$. Then, deeper features with reduced resolution are extracted and aggregated with the shallow feature

$$F_i^2 = \text{ResBlock}(\text{ResBlock}_{\downarrow}(F_i^1) \circ \text{MaxPool}_{\downarrow}(F_i^1)), \quad (9)$$

where $\text{ResBlock}(\cdot)$ and $\text{MaxPool}(\cdot)$ denote the residual block and max pooling, respectively; " $\downarrow$ " means that the resolution is decreased by half; " $\circ$ " denotes concatenation along the channel dimension. The third feature F_i^3 is obtained using the same operations. In this way, the cross-scale information is integrated in a top-to-bottom fashion. Besides, the information is also integrated in a bottom-to-top direction. Specifically, the spatial size of F_i^3 is increased by a factor of two and then fused with F_i^2 by

$$F_i^2 = \text{Conv}(F_i^2 \circ \text{Up}(F_i^3)), \quad (10)$$

where $\text{Up}(\cdot)$ denotes the nearest interpolation operation. The cross-scale information is also integrated into F_i^1 using the same methods.

Color Decoupling. Fig. 2(c) illustrates the flowchart for color decoupling. For $\forall i \in \{src, tar\}$ and $\forall j \in \{1, 2, 3\}$, the multiscale feature F_i^j is processed through convolutional layers to construct the color representation $F_{color}^{j,i}$ and the color-invariant feature $F_{invar}^{j,i}$. Inspired by [27], we facilitate this process with two loss functions. The first loss function focuses on color reconstruction and is defined as

$$L_{color}^{j,i} = \text{MSE}(\text{Net}_c(F_{color}^{j,i}), \text{Hist}(I_i)), \quad (11)$$

where $\text{Hist}(\cdot)$ is a color histogram function; $\text{Net}_c(\cdot)$ is a network that reconstructs the color information from the latent feature; $\text{MSE}(\cdot, \cdot)$ denotes the L2 loss. The second loss function addresses color decoupling and is computed as

$$L_{dis}^{j,i} = \|\text{CosSim}(F_{color}^{j,i}, F_{invar}^{j,i})\|_1, \quad (12)$$

where "$\|\cdot\|_1$" and $\text{CosSim}(\cdot, \cdot)$ represent the L1 loss and cosine similarity function, respectively. Minimizing this loss function encourages the color-invariant feature $F_{invar}^{j,i}$ to be orthogonal to the color feature $F_{color}^{j,i}$. Their correlations are minimized to decouple the color information from $F_{invar}^{j,i}$.

Homography Estimation. Fig. 2(d) illustrates the iterative homography estimation module. It utilizes the color-invariant features to estimate the offsets at two levels. For

Figure 3. Examples of synthetic data. The first row presents the synthesis results with various content weights. The second row shows the results with different template images. The third row displays the examples with different smoothing weights. The weights become smaller and greater from left to right for the first and third rows, respectively.

$\forall j \in \{1,2,3\}$, the estimation result derived from $F_{invar}^{src,j}$ and $F_{invar}^{tar,j}$ can be expressed as

$$O_{pred}^{j} = O_{pred}^{j,K} + O_{pred}^{j+1}, \tag{13}$$

where $O_{pred}^{j,K}$ is the residual offset output by the estimation block after K iterations and O_{pred}^{4} is initialized as a zero matrix. The first and second levels respectively utilize O_{pred}^{j+1} and $O_{pred}^{j,k-1}$ to assist in estimating $O_{pred}^{j,k}$ for $\forall k \in \{1,2,...,K\}$. $O_{pred}^{j,0}$ is also set to a zero matrix. The color-invariant features $F_{invar}^{src,j}$ and $F_{invar}^{tar,j}$ are used throughout the iterations. For simplicity, we denote F_{src}^{j} and F_{tar}^{j} as $F_{invar}^{src,j}$ and $F_{invar}^{tar,j}$, respectively. To derive $O_{pred}^{j,k}$, F_{src}^{j} is first warped by

$$F_{src}^{j,k} = \mathrm{Warp}(F_{src}^{j}, O_{pred}^{j,k-1} + O_{pred}^{j+1}), \tag{14}$$

Then, the similarity between the target and warped source image features is calculated as follows:

$$\begin{aligned} &C^{j,k}(u,v,m,n) \\ &= \sum_{u=-r}^{r}\sum_{v=-r}^{r} {F_{src}^{j,k}}^{T}(m+u,n+v)\cdot F_{tar}^{j}(m,n), \end{aligned} \tag{15}$$

where (m,n) denotes the coordinates of a feature vector; "T" represents the transpose operation; u and v enable the capture of contextual similarities; r is the searching radius. The resulting 4D tensor $C^{j,k}$ encodes the contextual correlations between the target and source images. $C^{j,k}$ is reshaped into the size of $(2\cdot r+1)^2 \times S/2^{j-1} \times S/2^{j-1}$ and input into the estimation block to update the residual offset

$$O_{pred}^{j,k} = \mathrm{Net}_h(C^{j,k}) + O_{pred}^{j,k-1}, \tag{16}$$

where $\mathrm{Net}_h(\cdot)$ comprises convolutional layers. O_{pred}^{1} is the final output of the iterative homography estimation module.

3.4. Loss Function

Since the synthetic data includes the ground-truth offsets, the model is trained in a supervised manner. The loss function is defined as

$$L = L_{pred} + \lambda \cdot \sum_{i\in\{src,tar\}}\sum_{j=1}^{3}\left(L_{color}^{j,i} + L_{dis}^{j,i}\right), \tag{17}$$

$$L_{pred} = \sum_{O_{pred}\in\mathcal{O}} \|O_{pred} - O_{gt}\|_1, \tag{18}$$

where λ is a hyperparameter; $\mathcal{O}$ represents the set $\{O_{pred}^{j,k} + O_{pred}^{j+1} | j \in \{1,2,3\}, k \in \{1,2,...,K\}\}$; The terms $L_{color}^{j,i}$ and $L_{dis}^{j,i}$ are defined in Eqs. (11) and (12), respectively. The term L_{pred} encourages the predicted offsets to approximate the ground truth as closely as possible. The other two terms in Eq. (17) enforce the decoupling of color information from the features utilized for homography estimation.

4. Experiments

4.1. Experimental Settings

Datasets. The proposed training data synthesis method and the homography estimation network were evaluated across four datasets: GoogleMap [45], GoogleEarth [45], RGB-NIR [3], and PDSCOCO [17]. The GoogleMap dataset includes paired images featuring both Google Maps and satellite map styles. The GoogleEarth dataset consists of images captured in Greater Boston throughout various seasons. The RGB-NIR dataset contains pairs of an RGB image and a

Table 1. Cross-dataset evaluation for GoogleMap. "$*$" means that the model is trained on our synthetic data. "$+$" indicates that the data synthesis method is applied to GoogleMap for training.

Method	GoogleEarth	RGB-NIR	PDSCOCO
DHN [12]	27.653	26.605	29.856
DHN*	22.933	13.864	14.173
MHN [18]	39.474	30.372	33.490
MHN*	3.110	7.549	4.251
IHN [5]	3.038	12.491	5.352
IHN$^+$	2.770	7.456	4.481
IHN*	1.853	5.647	1.684
MCNet [49]	20.518	16.557	8.202
MCNet$^+$	4.198	10.450	8.489
MCNet*	1.402	5.239	1.423
SCPNet [43]	5.508	10.918	4.862
AltO [30]	9.089	18.651	12.398
SSHNet [41]	24.952	30.535	25.486

Table 2. Cross-dataset evaluation for GoogleEarth. "$*$" means that the model is trained on our synthetic data. "$+$" indicates that the data synthesis method is applied to GoogleEarth for training.

Method	GoogleMap	RGB-NIR	PDSCOCO
DHN [12]	24.604	21.937	21.348
DHN*	11.321	13.864	14.173
MHN [18]	26.806	20.581	11.518
MHN*	11.218	7.549	4.251
IHN [5]	14.962	12.616	5.278
IHN$^+$	7.223	5.647	1.684
IHN*	5.303	8.336	2.449
MCNet [49]	10.799	9.982	3.753
MCNet$^+$	6.334	7.048	1.607
MCNet*	5.093	5.239	1.423
SCPNet [43]	28.291	14.281	4.460
AltO [30]	17.835	17.306	9.566
SSHNet [41]	21.335	17.419	10.114

Table 3. Cross-dataset evaluation for RGB-NIR. "$*$" means that the model is trained on our synthetic data. "$+$" indicates that the data synthesis method is applied to RGB-NIR for training.

Method	GoogleMap	GoogleEarth	PDSCOCO
DHN [12]	15.188	18.288	19.962
DHN*	11.321	22.933	14.173
MHN [18]	16.931	4.265	4.927
MHN*	11.218	3.110	4.251
IHN [5]	10.629	2.095	2.873
IHN$^+$	3.768	1.894	1.901
IHN*	5.303	1.853	1.684
MCNet [49]	6.312	1.843	1.451
MCNet$^+$	3.605	1.569	2.245
MCNet*	5.093	1.402	1.423
SCPNet [43]	19.109	10.618	3.749
AltO [30]	15.756	5.181	7.760
SSHNet [41]	23.680	2.848	3.879

near-infrared image. PDSCOCO features images that differ in brightness, saturation, contrast, and hue.

Implementations. Our training data synthesis method and homography estimation network are implemented using PyTorch. The training is conducted on a single RTX A6000 GPU. The AdamW is employed for parameter updates. For data synthesis, IEContraAST [8] is employed as the style transfer network. Content images are sourced from the MSCOCO dataset [21] to enable zero-shot estimation. Template images are sampled from the Painter by Numbers dataset [29]. The training iterations, learning rate, β, λ, batch size, K, and r are set to 1.2×10^5, 4×10^{-4}, 1×10^{-3}, 0.5, 16, 2, and 4, respectively.

Baselines. The baselines encompass supervised and unsupervised methods. The supervised methods include DHN [12], MHN [18], IHN [5], and MCNet [49]. The unsupervised methods consist of UDHN [23], CA-UDHN [42], SCPNet [43], AltO [30], and SSHNet [41].

Metrics. Following prior work, the mean average corner error (MACE) is adopted to evaluate the accuracy of homography estimation. A smaller MACE value indicates higher estimation accuracy.

4.2. Training Data Synthesis

Visualization. Fig. 3 showcases the results of rendering the same image with various parameters. As the content weights decrease, the images exhibit greater divergence from the original in terms of texture and color. Applying different styles can result in a variety of textures and colors. Additionally, a larger smoothing weight yields smoother textures. Despite this diversity, the structural information of the original image is preserved. This combination of structural consistency and appearance diversity enhances the model's ability to generalize across various modalities.

Cross-Dataset Evaluation. We first assess the effectiveness of the proposed data synthesis method through cross-dataset evaluation. Higher accuracy signifies better generalization performance. Tabs. 1 to 4 present the results for the baselines trained on GoogleMap, GoogleEarth, RGB-NIR, and PDSCOCO. These tables also include results for training on our synthetic data, which enables zero-shot estimation. The findings indicate that the baselines exhibit unsatisfactory generalization performance when trained on existing datasets, particularly with GoogleMap, GoogleEarth, and PDSCOCO. In contrast, the RGB-NIR dataset affords improved generalization. The near-infrared wavelength captures less texture and color information, allowing the structural information to dominate the images. Training on our synthetic data enhances the generalization perfor-

Table 4. Cross-dataset evaluation for PDSCOCO. "$*$" means that the model is trained on our synthetic data. "$+$" indicates that the data synthesis method is applied to PDSCOCO for training.

Method	GoogleMap	GoogleEarth	RGB-NIR
DHN [12]	18.315	11.961	12.669
DHN*	11.321	22.933	13.864
MHN [18]	22.822	5.694	11.272
MHN*	11.218	3.110	7.549
IHN [5]	21.727	2.156	10.745
IHN*	5.303	1.853	5.647
MCNet [49]	22.276	2.083	9.186
MCNet*	5.093	1.402	5.239
SCPNet [43]	6.553	11.673	12.528
AltO [30]	27.060	5.945	14.739
SSHNet [41]	23.586	2.105	11.010

mance of the baselines. The generalization performance of zero-shot DHN declines in three cases due to its weaker learning capability compared to other methods. Excluding these cases, improvements facilitated by our data synthesis method range from 1.93% to 93.17%, with over 50% improvement observed in roughly half of the cases. Besides, PDSCOCO is also derived from MSCOCO and exhibits variations in brightness, saturation, hue, and contrast. The generalization performance remains unsatisfactory when training models on it. This suggests that our synthesis strategy contributes to better generalization beyond the variability offered by the content images in MSCOCO.

Augmentation. We regard the synthesis method as a form of augmentation and apply it to existing datasets for further evaluation. IHN and MCNet are selected for testing due to their superior estimation capabilities. The results are presented in Tabs. 1 to 3. Although estimation accuracy decreases in two cases, the synthesis method generally enhances the generalization performance by 8.82% to 79.54%. This indicates that our synthesis strategy is beneficial for generalization, irrespective of the content images used.

Within-Dataset Evaluation. Tab. 5 reports the within-dataset and zero-shot accuracy of the baselines. Both supervised and unsupervised methods are trained and tested on the same dataset, while the zero-shot methods are trained on data synthesized from MSCOCO. As observed, the supervised methods generally achieve the highest within-dataset performance, as they leverage ground-truth offsets to adapt to the specific modalities of the training dataset. Applying our synthesis method to existing datasets enhances generalization but also results in a decrease in accuracy. IHN and MCNet are dissuaded by the synthetic data from fully utilizing modality information to infer the offsets. Furthermore, our synthetic data enhances zero-shot estimation due to its diversity in textures and colors. Conversely, the lack of modality information results in lower accuracy for within-dataset evaluations compared to supervised methods. Notably, the zero-shot baselines exhibit estimation performance comparable to that of unsupervised methods in within-dataset evaluations while demonstrating superior generalization capabilities. The unsupervised method SSHNet achieves significantly higher accuracy on GoogleMap than the zero-shot methods, as it employs image translation techniques to unify the modalities of image pairs. However, as shown in Tab. 1, this approach also substantially degrades generalization performance.

The supplementary material reports more experimental results for the training data synthesis method.

4.3. Cross-Scale and Color-Invariant Network

Quantitative Results. We compare our CCNet with the baselines when they are trained on the four datasets and our synthetic data. Tab. 5 shows the comparison results. As can be observed, our CCNet outperforms the baselines in both within-dataset and zero-shot estimation tasks. In the within-dataset evaluation, CCNet achieves higher accuracy than the second-best method, with improvements of 29.50%, 8.83%, 7.25%, and 5.74% on GoogleMap, GoogleEarth, RGB-NIR, and PDSCOCO, respectively. These enhancements underscore CCNet's superior capability in multimodal homography estimation. For zero-shot evaluation, CCNet shows improvements of 13.94%, 0.21%, 14.85%, and 3.87% on the four datasets. CCNet effectively integrates the cross-scale information into the features extracted for homography estimation, leading to enhanced results. Additionally, by decoupling color from the features, CCNet minimizes the negative effects associated with color information when processing multimodal image pairs.

Qualitative Results. Fig. 4 presents the visual results of within-dataset evaluation on GoogleEarth and RGB-NIR, while Fig. 5 showcases the visualizations for zero-shot evaluation on GoogleMap. As demonstrated, our CCFNet estimates the offsets more accurately than the baselines across both within-dataset and zero-shot evaluations.

Computational Costs. Tab. 6 presents the runtime and model size of various supervised methods. As indicated in Tab. 5 and Tab. 6, our CCFNet demonstrates superior homography estimation and generalization capabilities, requiring only a slight increase in runtime and parameters.

The supplementary material provides additional experimental results for CCNet.

5. Conclusion

In this paper, we introduce a training data synthesis method aimed at enhancing generalization in multimodal homography estimation. This method produces various unaligned image pairs by rendering the same image with different textures and colors, along with their corresponding ground-truth offsets. While these images exhibit diverse appear-

Table 5. Within-dataset and zero-shot evaluation results (MACE ↓) of our training data synthesis method and the homography estimation network. The best result is written in bold while the second-best one is underlined. "+" indicates that the models are trained on the dataset augmented by our synthesis method. "Zero-shot" means that the models are trained on our synthetic method.

Type	Method	GoogleMap	GoogleEarth	RGB-NIR	PDSCOCO
Supervised	DHN [12]	3.556	7.735	12.855	11.937
	MHN [18]	1.561	2.061	7.078	5.142
	IHN [5]	1.025	1.147	4.588	1.445
	MCNet [49]	<u>0.261</u>	<u>0.577</u>	<u>3.226</u>	<u>1.062</u>
	CCNet (Ours)	**0.184**	**0.526**	**2.992**	**1.001**
	IHN$^+$	1.515	1.236	3.552	1.498
	MCNet$^+$	0.992	0.776	3.146	1.348
Unupervised	UDHN [23]	24.713	21.248	25.452	25.684
	CA-UDHN [42]	24.557	23.761	24.297	24.875
	SCPNet [43]	4.364	2.794	10.618	9.448
	AltO [30]	4.739	3.174	7.897	4.979
	SSHNet [41]	1.394	5.888	6.743	1.610
Zero-shot	DHN [12]	11.321	22.933	13.864	14.173
	MHN [18]	11.218	3.110	7.549	4.251
	IHN [5]	5.303	1.853	5.647	1.684
	MCNet [49]	5.093	1.402	5.239	1.423
	CCNet (Ours)	**4.383**	**1.399**	**4.461**	**1.368**

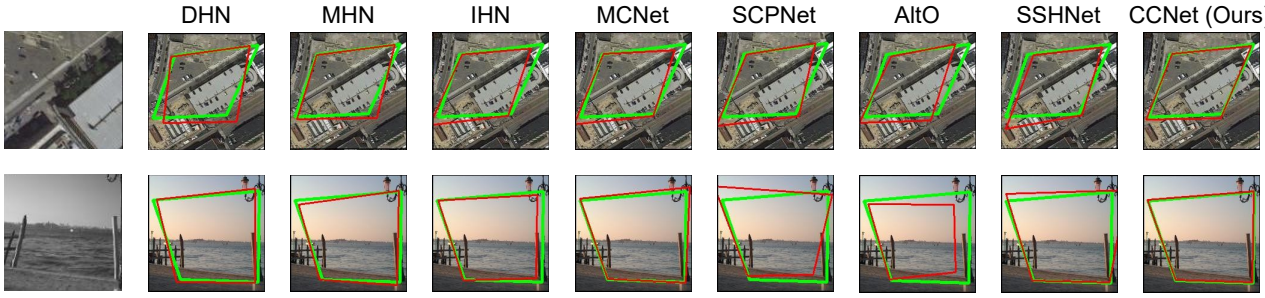

Figure 4. Visualization results of within-dataset evaluation. The first row presents the estimation results from GoogleEarth, while the second row features results from RGB-NIR. The first column displays the source image to be warped, while the other columns demonstrate the target images. Greater similarity between the red and green quadrilaterals indicates higher accuracy in the estimation.

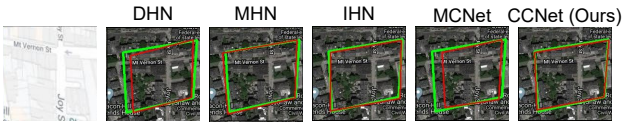

Figure 5. Visualization results of zero-shot evaluation using testing images from GoogleMap. The first column displays the source image to be warped. A greater similarity between the red and green quadrilaterals indicates higher accuracy in estimation.

ances, they retain the structural information of the original image. Therefore, a homography estimation model can be trained on these data in a supervised manner to achieve zero-shot estimation. Additionally, the synthesis strategy can be applied to existing datasets to improve generalization, albeit with a slight compromise in within-dataset performance. We also design a network to enhance homography estimation accuracy. This network fully leverages cross-scale information to enhance the estimation results. Color information is decoupled from the features to boost the network's ability in multimodal image processing. Extensive experiments are conducted to demonstrate the effectiveness of the proposed data synthesis method and the superiority of our homography estimation network.

Table 6. Computational costs of supervised methods and our network. The first row displays the runtime (in milliseconds) while the second row presents the model size (in megabytes).

DHN [12]	MHN [18]	IHN [5]	MCNet [49]	CCNet
11.21	10.92	26.14	31.01	32.73
34.20	2.57	0.76	0.85	1.21

Acknowledgements

This work was funded in part by the Science and Technology Development Fund, Macau SAR (File no. 0050/2024/AGJ), by the University of Macau and University of Macau Development Foundation (File no. MYRG-GRG2024-00181-FST-UMDF).

References

[1] Moab Arar, Yiftach Ginger, Dov Danon, Amit H Bermano, and Daniel Cohen-Or. Unsupervised multi-modal image registration via geometry preserving image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13410–13419, 2020. 1

[2] Simon Baker and Iain Matthews. Lucas-kanade 20 years on: A unifying framework. *International journal of computer vision*, 56(3):221–255, 2004. 1

[3] Matthew Brown and Sabine Süsstrunk. Multi-spectral sift for scene category recognition. In *CVPR 2011*, pages 177–184. IEEE, 2011. 5

[4] Wenxiao Cai and Wankou Yang. Object-level geometric structure preserving for natural image stitching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1926–1934, 2025. 1

[5] Si-Yuan Cao, Jianxin Hu, Zehua Sheng, and Hui-Liang Shen. Iterative deep homography estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1879–1888, 2022. 1, 2, 6, 7, 8

[6] Si-Yuan Cao, Runmin Zhang, Lun Luo, Beinan Yu, Zehua Sheng, Junwei Li, and Hui-Liang Shen. Recurrent homography estimation using homography-guided image warping and focus transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9833–9842, 2023. 2

[7] 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. 2

[8] Haibo Chen, lei zhao, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. Artistic style transfer with internal-external learning and contrastive learning. 34:26561–26573, 2021. 3, 6

[9] Jiaxin Cheng, Ayush Jaiswal, Yue Wu, Pradeep Natarajan, and Prem Natarajan. Style-aware normalized loss for improving arbitrary style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 134–143, 2021. 3

[10] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8795–8805, 2024. 3

[11] Zhipeng Deng, Luyang Luo, and Hao Chen. Enable the right to be forgotten with federated client unlearning in medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 240–250. Springer, 2024. 1

[12] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Deep image homography estimation. *arXiv preprint arXiv:1606.03798*, 2016. 1, 2, 6, 7, 8

[13] Elan Dubrofsky. Homography estimation. *Diplomová práce. Vancouver: Univerzita Britské Kolumbie*, 5, 2009. 1

[14] 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. 2

[15] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 3

[16] Seungryong Kim, Dongbo Min, Bumsub Ham, Seungchul Ryu, Minh N Do, and Kwanghoon Sohn. DASC: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2103–2112, 2015. 1

[17] Daniel Koguciuk, Elahe Arani, and Bahram Zonooz. Perceptual loss for robust unsupervised homography estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4274–4283, 2021. 5

[18] 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. 2, 6, 7, 8

[19] Huafeng Li, Zengyi Yang, Yafei Zhang, Wei Jia, Zhengtao Yu, and Yu Liu. MulFS-CAP: Multimodal fusion-supervised cross-modality alignment perception for unregistered infrared-visible image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1

[20] Jiaxue Li and Yicong Zhou. Automatic quaternion-domain color image stitching. *IEEE Transactions on Image Processing*, 33:1299–1312, 2024. 1

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[22] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 1

[23] Ty Nguyen, Steven W Chen, Shreyas S Shivakumar, Camillo Jose Taylor, and Vijay Kumar. Unsupervised deep homography: A fast and robust homography estimation model. *IEEE Robotics and Automation Letters*, 3(3):2346–2353, 2018. 1, 2, 6, 8

[24] 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. 1

[25] Lang Nie, Chunyu Lin, Kang Liao, Shuaicheng Liu, and Yao Zhao. Parallax-tolerant unsupervised deep image stitching.

In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7365–7374, 2023. 1

[26] Fu-Zhao Ou, Chongyi Li, Shiqi Wang, and Sam Kwong. MR-FIQA: Face image quality assessment with multi-reference representations from synthetic data generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12915–12925, 2025. 1

[27] Priyank Pathak and Yogesh S Rawat. Colors see colors ignore: Clothes changing reid with color disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16797–16807, 2025. 2, 4

[28] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to SIFT or SURF. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 1

[29] small yellow duck and Wendy Kan. Painter by Numbers. Available at https://www.kaggle.com/c/painter-by-numbers. 6

[30] Sanghyeob Song, Jaihyun Lew, Hyemi Jang, and Sungroh Yoon. Unsupervised homography estimation on multimodal image pair via alternating optimization. *Advances in Neural Information Processing Systems*, 37:61306–61327, 2024. 1, 2, 6, 7, 8

[31] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 3508–3515. International Joint Conferences on Artificial Intelligence Organization, 2022. 1

[32] Xiaohang Wang, Xuanhong Chen, Bingbing Ni, Zhengyan Tong, and Hang Wang. Learning continuous depth representation via geometric spatial aggregator. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2698–2706, 2023. 1

[33] Xiao Wu, Zi-Han Cao, Ting-Zhu Huang, Liang-Jian Deng, Jocelyn Chanussot, and Gemine Vivone. Fully-connected transformer for multi-source image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3): 2071–2088, 2025. 1

[34] Han Xu, Jiayi Ma, Jiteng Yuan, Zhuliang Le, and Wei Liu. RFNet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19679–19688, 2022. 1

[35] Li Xu, Cewu Lu, Yi Xu, and Jiaya Jia. Image smoothing via L_0 gradient minimization. In *Proceedings of the 2011 SIGGRAPH Asia conference*, pages 1–12, 2011. 3

[36] Nianjin Ye, Chuan Wang, Haoqiang Fan, and Shuaicheng Liu. Motion basis learning for unsupervised deep homography estimation with subspace projection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13117–13125, 2021. 1, 2

[37] Xinchen Ye, Yanjun Guo, Baoli Sun, Rui Xu, Zhihui Wang, and Haojie Li. C^2ANet: Cross-scale and cross-modality aggregation network for scene depth super-resolution. *IEEE Transactions on Multimedia*, 26:2574–2584, 2024. 1

[38] Jinkun You and Yicong Zhou. Two-stage watermark removal framework for spread spectrum watermarking. *IEEE Transactions on Multimedia*, 26:7687–7699, 2024. 1

[39] Jinkun You, Jie Zhang, and Yicong Zhou. HFFNet: Hierarchical feature fusion network for thermal image super-resolution. *Available at SSRN 5372220*. 1

[40] Jinkun You, Jiaxue Li, Jie Zhang, and Yicong Zhou. Dense cross-scale image alignment with fully spatial correlation and just noticeable difference guidance. *arXiv preprint arXiv:2511.09028*, 2025. 2

[41] Junchen Yu, Si-Yuan Cao, Runmin Zhang, Chenghao Zhang, Zhu Yu, Shujie Chen, Bailin Yang, and Hui-Liang Shen. SSHNet: Unsupervised cross-modal homography estimation via problem reformulation and split optimization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16685–16694, 2025. 1, 2, 6, 7, 8

[42] Jirong Zhang, Chuan Wang, Shuaicheng Liu, Lanpeng Jia, Nianjin Ye, Jue Wang, Ji Zhou, and Jian Sun. Content-aware unsupervised deep homography estimation. In *European conference on computer vision*, pages 653–669. Springer, 2020. 1, 2, 6, 8

[43] Runmin Zhang, Jun Ma, Si-Yuan Cao, Lun Luo, Beinan Yu, Shu-Jie Chen, Junwei Li, and Hui-Liang Shen. Scpnet: Unsupervised cross-modal homography estimation via intra-modal self-supervised learning. In *Computer Vision – ECCV 2024*, pages 460–477, Cham, 2025. Springer Nature Switzerland. 1, 2, 6, 7, 8

[44] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10146–10156, 2023. 3

[45] 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. 1, 2, 5

[46] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5906–5916, 2023. 1

[47] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25912–25921, 2024. 1

[48] Qiang Zhou and Xin Li. STN-homography: Direct estimation of homography parameters for image pairs. *Applied Sciences*, 9(23):5187, 2019. 2

[49] Haokai Zhu, Si-Yuan Cao, Jianxin Hu, Sitong Zuo, Beinan Yu, Jiacheng Ying, Junwei Li, and Hui-Liang Shen. MCNet: Rethinking the core ingredients for accurate and efficient homography estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25932–25941, 2024. 2, 6, 7, 8