

# SSHNet: Unsupervised Cross-modal Homography Estimation via Problem Reformulation and Split Optimization

Junchen Yu<sup>1,2</sup> Si-Yuan Cao<sup>1,2,3\*</sup> Runmin Zhang<sup>2</sup> Chenghao Zhang<sup>2</sup>  
Zhu Yu<sup>2</sup> Shujie Chen<sup>4</sup> Bailin Yang<sup>4</sup> Hui-Liang Shen<sup>2</sup>

<sup>1</sup>Ningbo Innovation Center, Zhejiang University <sup>2</sup>College of Information Science and Electronic Engineering, Zhejiang University

<sup>3</sup>NingboTech University <sup>4</sup>Zhejiang Key Laboratory of Big Data and Future ECommerce Technology, Hangzhou, China

{yujunchen, cao\_siyuan, runmin\_zhang, zch00, yu\_zhu}@zju.edu.cn

{chenshujie, ybl}@zjgsu.edu.cn shenhl@zju.edu.cn

## 1. Implementation Details

We implement our SSHNet using PyTorch, and train it with the AdamW [12] optimizer. The learning rate scheduler adopts OneCycleLR [16], with a maximum learning rate set to  $3 \times 10^{-4}$ . The training is conducted with a batch size of 16 for 120,000 iterations. All the experiments are conducted on a single NVIDIA RTX4090 GPU.

## 2. Parameterization of Homography Matrix

Following previous approaches [3–5, 9, 19], we parameterize the homography matrix using the displacement vectors of the four corner points. The homography matrix can be obtained by solving the least squares problem,

$$\mathbf{A}\mathbf{h} = \mathbf{b}, \quad (1)$$

where  $\mathbf{b}$  is the coordinates of the warped four corner points,  $\mathbf{A}$  is composed of the warped four corner points and the original four corner points,  $\mathbf{h}$  is the vectorized homography matrix, which is formulated as

$$\mathbf{h} = [\mathbf{H}_{11} \ \mathbf{H}_{12} \ \mathbf{H}_{13} \ \mathbf{H}_{21} \ \mathbf{H}_{22} \ \mathbf{H}_{23} \ \mathbf{H}_{31} \ \mathbf{H}_{32}]^\top. \quad (2)$$

For a corner point  $x = (u, v)$  in the source image  $\mathbf{I}_A$ , its corresponding point  $x' = (u', v')$  in the target image  $\mathbf{I}_B$  can be formulated as

$$\begin{aligned} u' &= \frac{\mathbf{H}_{11}u + \mathbf{H}_{12}v + \mathbf{H}_{13}}{\mathbf{H}_{31}u + \mathbf{H}_{32}v + 1} \\ v' &= \frac{\mathbf{H}_{21}u + \mathbf{H}_{22}v + \mathbf{H}_{23}}{\mathbf{H}_{31}u + \mathbf{H}_{32}v + 1}. \end{aligned} \quad (3)$$

When the deformation of four corner points are known, the multivariate equation for the elements of  $\mathbf{h}$  can be solved using a least squares approach. We define the four corner

points in  $\mathbf{I}_A$  as  $(u_i, v_i)$ , in  $\mathbf{I}_B$  as  $(u'_i, v'_i)$ , where  $i = 1, 2, 3, 4$ .  $\mathbf{A}$  is formulated as

$$\mathbf{A} = \begin{bmatrix} u_1 & v_1 & 1 & 0 & 0 & 0 & -u_1u'_1 & -v_1u'_1 \\ 0 & 0 & 0 & u_1 & v_1 & 1 & -u_1v'_1 & -v_1v'_1 \\ u_2 & v_2 & 1 & 0 & 0 & 0 & -u_2u'_2 & -v_2u'_2 \\ 0 & 0 & 0 & u_2 & v_2 & 1 & -u_2v'_2 & -v_2v'_2 \\ u_3 & v_3 & 1 & 0 & 0 & 0 & -u_3u'_3 & -v_3u'_3 \\ 0 & 0 & 0 & u_3 & v_3 & 1 & -u_3v'_3 & -v_3v'_3 \\ u_4 & v_4 & 1 & 0 & 0 & 0 & -u_4u'_4 & -v_4u'_4 \\ 0 & 0 & 0 & u_4 & v_4 & 1 & -u_4v'_4 & -v_4v'_4 \end{bmatrix}, \quad (4)$$

and  $\mathbf{b}$  as

$$\mathbf{b} = [u'_1 \ v'_1 \ u'_2 \ v'_2 \ u'_3 \ v'_3 \ u'_4 \ v'_4]^\top. \quad (5)$$

Then the predicted deformation cube  $\mathbf{P}$  can be expressed as

$$\begin{aligned} u'_1 &= u_1 + \mathbf{P}(0, 0, 0), \\ v'_1 &= v_1 + \mathbf{P}(1, 0, 0), \\ u'_2 &= u_2 + \mathbf{P}(0, 0, 1), \\ v'_2 &= v_2 + \mathbf{P}(1, 0, 1), \\ u'_3 &= u_3 + \mathbf{P}(0, 1, 0), \\ v'_3 &= v_3 + \mathbf{P}(1, 1, 0), \\ u'_4 &= u_4 + \mathbf{P}(0, 1, 1), \\ v'_4 &= v_4 + \mathbf{P}(1, 1, 1). \end{aligned} \quad (6)$$

## 3. Details of Modality Transfer Network

Inspired by Swin-Unet [2], we design a transformer-based U-shaped generator as the modality transfer network, which consists of encoder, bottleneck, decoder and skip connections. We illustrate the architecture of the modality transfer network in Fig. 1. The input image  $\mathbf{I}$  is initially projected into feature space using a  $3 \times 3$  convolutional layer. The encoder then processes the feature using 2 Swin Transformer blocks, a PixelShuffle layer, and an  $1 \times 1$  convolutional layer

\*Corresponding author.

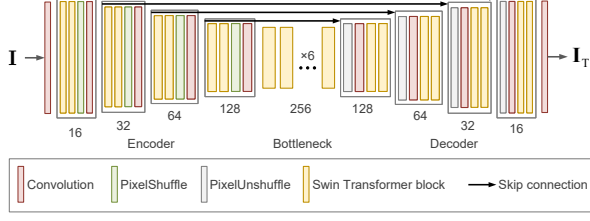


Figure 1. The detailed architecture of the modality transfer network.

to downsample the feature and expand its channel by  $2\times$ . Correspondingly, the decoder upsamples the feature and reduces its channel by  $2\times$  using a PixelUnshuffle layer, an  $1\times 1$  convolutional layer, and 2 Swin Transformer blocks. Both two procedures are repeated four times in the modality transfer network. The bottleneck between the encoder and decoder is constructed with 6 Swin Transformer blocks to learn deep feature representations. The shallow and deep features are concatenated together via skip connections to complement the loss of spatial information caused by down-sampling. At the end of the modality transfer network, an  $1\times 1$  convolutional layer projects the output feature into 3 channels, producing the modality transfer result  $\mathbf{I}_T$ .

The basic unit of the modality transfer network is Swin Transformer block [11], different from the standard multi-head self-attention, Swin Transformer block first partitions the inputs into non-overlapping local windows, each window contains  $M\times M$  patches. It computes the local attention for each window:

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(QK^T/\sqrt{d}\right)V, \quad (7)$$

where  $Q, K, V \in \mathbb{R}^{M^2 \times d}$  denote the query, key and value matrices,  $M^2$  is the number of patches in a window,  $d$  is the query/key dimension.

#### 4. Implementation of Extra Homography Feature Space Supervision

As discussed in our main text, we introduce an extra homography feature space supervision, which is implemented during the optimization of Sub-problem II. While this supervision primarily targets the optimization of the modality transfer network, it is further extended to ensure that the homography estimation network can extract robust features from  $\mathbf{I}_B$  and the transferred image  $\mathbf{I}_{A,T}$ . To achieve this, we incorporate a correlation-based homography feature loss to simultaneously optimize the modality transfer network and the feature extractor of the homography estimation network in practice. Finally, the entire optimization process of Sub-problem II can be formulated as

$$\begin{aligned} \arg \min_{\theta, \xi} & L_T(\mathcal{T}_\theta(\mathbf{I}_A), \mathcal{W}(\mathbf{I}_B, \mathcal{H}_{\zeta^*}(\mathbf{I}_{A,T}, \mathbf{I}_B))) \\ & + L_{HF}(\mathcal{F}_\xi(\mathcal{T}_\theta(\mathbf{I}_A)), \mathcal{F}_\xi(\mathcal{W}(\mathbf{I}_B, \mathcal{H}_{\zeta^*}(\mathbf{I}_{A,T}, \mathbf{I}_B)))) \end{aligned} \quad (8)$$

where  $\mathcal{T}_\theta$  denotes the modality transfer network with parameters  $\theta$  to be optimized,  $\mathcal{F}_\xi$  denotes the feature extractor with parameters  $\xi$  to be optimized,  $\mathcal{H}_{\zeta^*}$  denotes the homography estimation network with parameters  $\zeta^*$  frozen,  $\mathcal{W}$  denotes the image warping with the estimated homography,  $L_T$  denotes the modality transfer loss, and  $L_{HF}$  denotes the homography feature loss.

#### 5. Details of Datasets

Figure 2 shows example image pairs from GoogleMap [20], DPDN [14], OPT-SAR [10], Flash/no-flash [7], and RGB/NIR [1] with offset of  $[-32, +32]$ . Following are details of each dataset.

**GoogleMap** is composed of aligned satellite and map images. We choose the satellite image as the source image and the map image as the target image. We then use the training and testing data shared in [20] with the size of  $192\times 192$ , including 8822 and 888 samples respectively. The  $128\times 128$  image pairs with homography deformation are produced by center cropping, enabling perturbation of offset  $[-32, +32]$ .

**DPDN** dataset contains RGB/Depth image pairs generated by a physically based renderer. We select RGB as the source and depth map as the target, and process the same way as GoogleMap. The training and testing splits contain 4000 and 1000 samples respectively.

**OPT-SAR** contains joint optical and synthetic aperture radar (SAR) images, initially used for land use classification. We choose optical image as the source, and SAR image as the target. Then we resize the image to  $192\times 192$  and generate image pairs the same way as GoogleMap. The training and testing splits contain 8000 and 2000 samples respectively.

**Flash/no-flash** contains 120 indoor and outdoor image pairs. We first resize the image to  $320\times 213$ , and then generate a  $128\times 128$  image pair with simulated homography in the same way as GoogleMap. The training and testing splits contain 60 and 60 samples respectively.

**RGB/NIR** dataset has images of the RGB and NIR spectral bands. We resize the image to  $256\times 256$ , then process the same way as GoogleMap. The training and testing splits contain 103 and 153 samples respectively.

#### 6. More Experimental Results

**Degree of the Deformations.** In real scenarios, the degree of deformation between the cross-modal image pairs is usually unknown. Therefore, we further alter the degree of the simulated deformations for self-supervised training and evaluate their effectiveness on different cross-modal homography deformations. We train our SSHNet with the simulated deformations of the range  $[-8, +8]$ ,  $[-16, +16]$ , and  $[-32, +32]$ , and evaluate the trained networks on the cross-modal deformations of the range  $[-8, +8]$ ,  $[-16, +16]$ , and  $[-32, +32]$ . The



Table 1. Ablation study of the degree of deformation. Simulated denotes the range of simulated deformation during self-supervised training, and Real denotes the range of cross-modal deformation during testing.

Real \ Simulated	Simulated		
	$[-8, +8]$	$[-16, +16]$	$[-32, +32]$
$[-8, +8]$	2.07	1.84	2.26
$[-16, +16]$	7.54	1.81	2.53
$[-32, +32]$	22.37	12.32	2.94

Table 2. Distillation training results.

Method \ Dataset	Dataset				
	GoogleMap	DPDN	OPT-SAR	Flash/no-flash	RGB/NIR
SSHNet-DHN	9.28	9.71	17.63	10.51	12.13
SSHNet-DHN-D	9.33	9.65	18.33	11.12	12.58
SSHNet-MHN	2.90	3.04	6.77	5.62	6.93
SSHNet-MHN-D	3.17	3.10	6.82	5.51	7.12
SSHNet-SCPNet	3.89	4.81	12.94	2.65	3.86
SSHNet-SCPNet-D	4.08	4.79	13.17	3.01	4.03
SSHNet-IHN	1.23	1.12	2.94	1.08	1.66
SSHNet-IHN-D	1.26	1.24	3.31	1.16	1.72
SSHNet-RHWF	1.29	1.26	3.08	0.95	1.52
SSHNet-RHWF-D	1.34	1.30	3.35	1.01	1.50

results are listed in Table 1. We find that the self-supervised training of SSHNet functions effectively when the simulated deformation is larger than the real cross-modal one. Therefore, we use a simulated deformation range of  $[-32, +32]$  for training, and recommend employing a relatively larger deformation degree for real-world scenarios.

**Distillation Training.** We conduct the distillation training on SSHNet with various homography estimation architectures, including DHN [5], MHN [9], SCPNet [19], IHN [3], and RHWF [4]. The experimental results of the distillation training are summarized in Table 2. Our experiments show that, regardless of the homography estimation architecture, the SSHNet-D obtained through distillation training consistently achieves performance comparable to that of the original SSHNet. This highlights the effectiveness and robustness of our distillation training technique.

**Training Stability.** As mentioned in our main text, our SSHNet demonstrates substantial training stability and can cooperate with various homography estimation architectures. To further illustrate the stability of our framework, we adopt two iterative homography estimation architectures, IHN and RHWF, in our SSHNet and the previous unsupervised SOTA approach SCPNet. The results in Table 3 show that SCPNet fails to cooperate with IHN and RHWF. The reliance on feature map content consistency for homography estimation hinders the convergence of SCPNet when using iterative homography estimation architectures. In contrast, our SSHNet relies solely on simulated deformation for direct supervision of homography estimation, leading to exceptional training stability. When adopting iterative homography architectures, our SSHNet can produce satisfactory homography estimation results.

Table 3. Comparison of SSHNet and SCPNet using iterative homography estimation architectures. NC denotes the training is not converged.

Method \ Dataset	Dataset				
	GoogleMap	DPDN	OPT-SAR	Flash/no-flash	RGB/NIR
SCPNet-IHN	NC	NC	NC	NC	NC
SCPNet-RHWF	NC	NC	NC	NC	NC
SSHNet-IHN	1.23	1.12	2.94	1.08	1.66
SSHNet-RHWF	1.29	1.26	3.08	0.95	1.52

**Modality Transfer Results.** We illustrate more visualization of modality transfer results on GoogleMap, DPDN, and OPT-SAR datasets in Figure 3. It can be seen that the modality transfer results yield good quality and preserve details of objects, which further improves homography estimation performance.

**Homography Estimation Results.** Figure 4 visualizes more homography estimation results of different approaches on each dataset. It is observed that UDHN [13], biHome [8], CA-UDHN [18], BasesHomo [17], and UMF-CMGR [6] all struggle to find reasonable results. The previous SOTA unsupervised approach, SCPNet, demonstrates higher accuracy on datasets like GoogleMap, Flash/no-Flash, and RGB/NIR. However, it fails to provide reliable predictions on the DPDN and OPT-SAR datasets. DHN, MHN, and LocalTrans [15] produces relatively more accurate homography estimations with direct supervision. Notably, our SSHNet-IHN achieves higher accuracy than SCPNet, DHN, MHN, and LocalTrans.

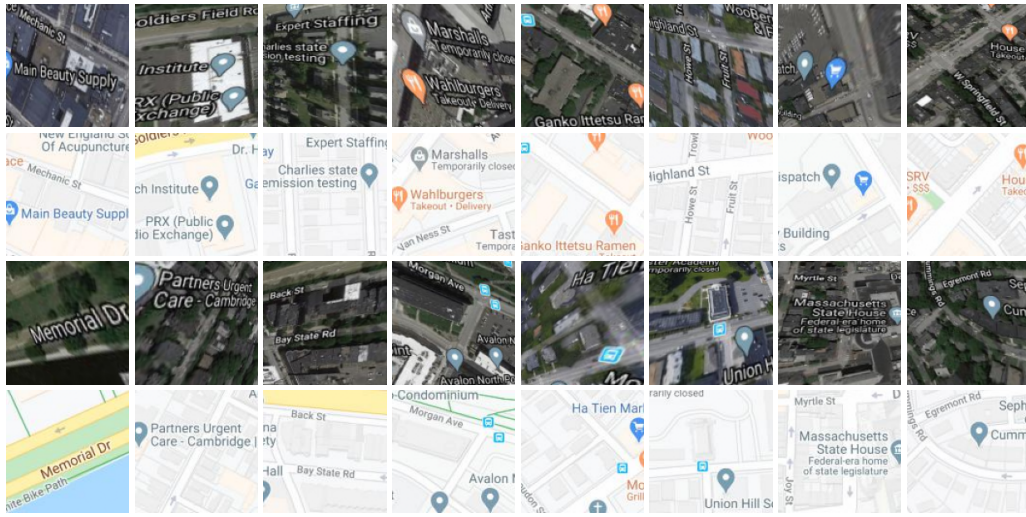
## 7. Limitation

Although the proposed SSHNet significantly outperforms other unsupervised methods, it requires the modality transfer network and the homography estimation network to be trained together. This results in increased GPU memory usage and longer training times.

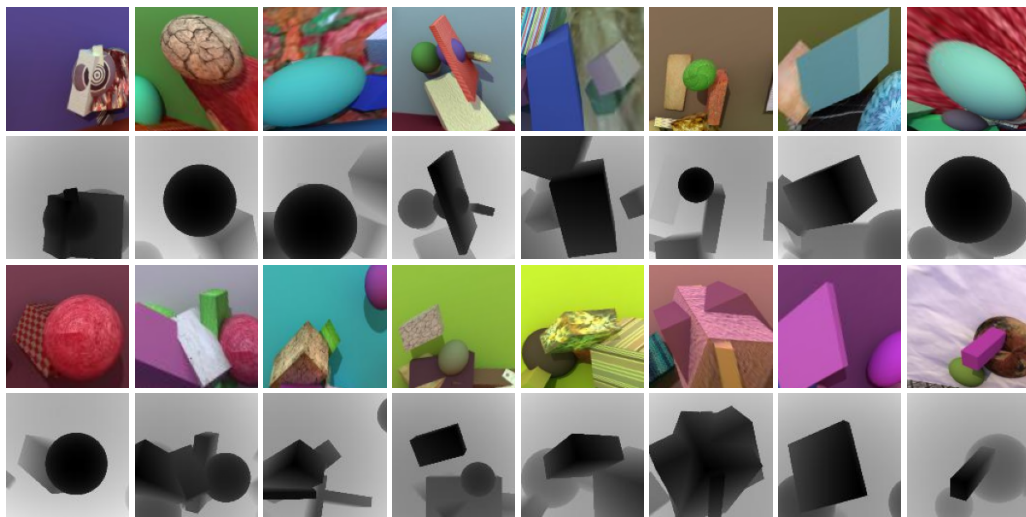
## References

- [1] Matthew Brown and Sabine Süsstrunk. Multispectral SIFT for scene category recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 177–184. IEEE, 2011. 2
- [2] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 205–218. Springer, 2022. 1
- [3] 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, 3
- [4] 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. 3

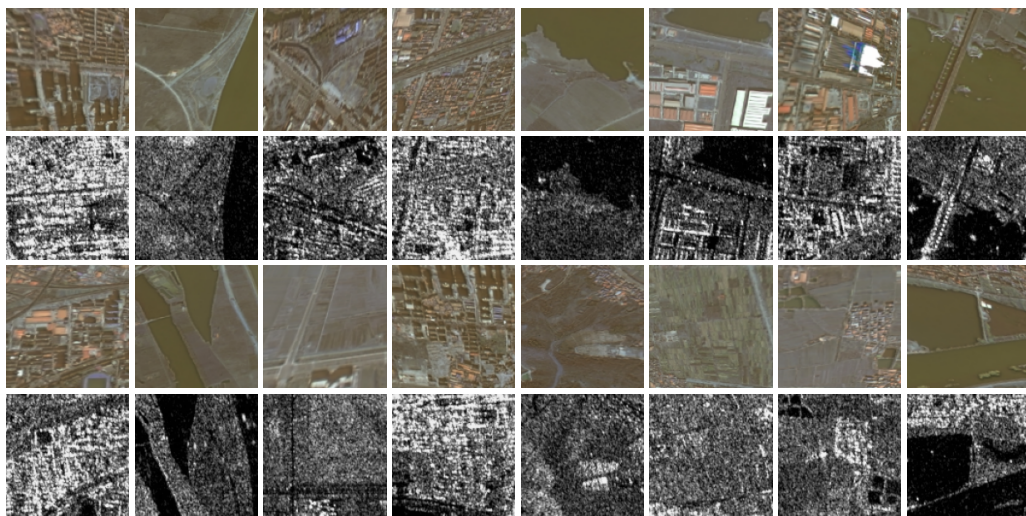
- [5] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Deep image homography estimation. *arXiv preprint arXiv:1606.03798*, 2016. [1](#), [3](#)
- [6] Wang Di, Liu Jinyuan, Fan Xin, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2022. [3](#)
- [7] Shengfeng He and Rynson WH Lau. Saliency detection with flash and no-flash image pairs. In *Proceedings of the European Conference on Computer Vision*, pages 110–124. Springer, 2014. [2](#)
- [8] 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. [3](#)
- [9] 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. [1](#), [3](#)
- [10] Xue Li, Guo Zhang, Hao Cui, Shasha Hou, Shun Yao Wang, Xin Li, Yujia Chen, Zhijiang Li, and Li Zhang. MCANet: A joint semantic segmentation framework of optical and SAR images for land use classification. *International Journal of Applied Earth Observation and Geoinformation*, 106:102638, 2022. [2](#)
- [11] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. [2](#)
- [12] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. [1](#)
- [13] 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. [3](#)
- [14] Gernot Riegler, David Ferstl, Matthias Rüther, and Horst Bischof. A deep primal-dual network for guided depth super-resolution. In *British Machine Vision Conference*. The British Machine Vision Association, 2016. [2](#)
- [15] 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. [3](#)
- [16] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, pages 369–386. SPIE, 2019. [1](#)
- [17] 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. [3](#)
- [18] Jirong Zhang, Chuan Wang, Shuaicheng Liu, Lanpeng Jia, Nianjin Ye, Jue Wang, Ji Zhou, and Jian Sun. Content-aware unsupervised deep homography estimation. In *Proceedings of the European Conference on Computer Vision*, pages 653–669. Springer, 2020. [3](#)
- [19] Runmin Zhang, Jun Ma, Si-Yuan Cao, Lun Luo, Beinan Yu, Shu-Jie Chen, Junwei Li, and Hui-Liang Shen. SCP-Net: Unsupervised cross-modal homography estimation via intra-modal self-supervised learning. In *Proceedings of the European Conference on Computer Vision*, pages 460–477. Springer, 2024. [1](#), [3](#)
- [20] 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. [2](#)



(a) GoogleMap



(b) DP3D



(c) OPT-SAR





(d) Flash/no-flash



(e) RGB/NIR

Figure 2. Example image pairs from GoogleMap, DPDN, OPT-SAR, Flash/no-flash, and RGB/NIR datasets, with offset of  $[-32, +32]$ .

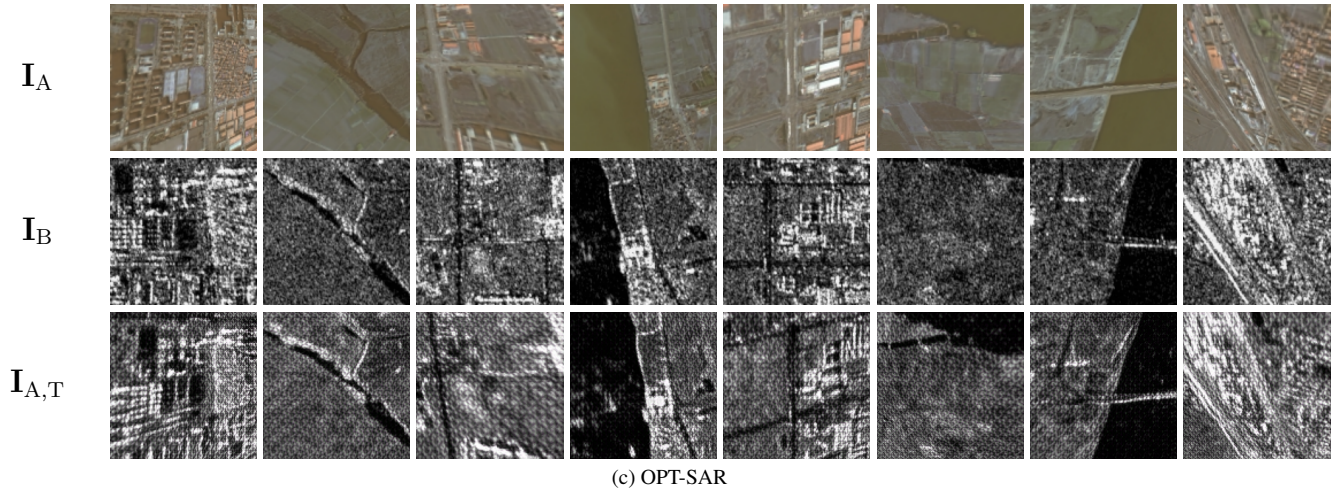
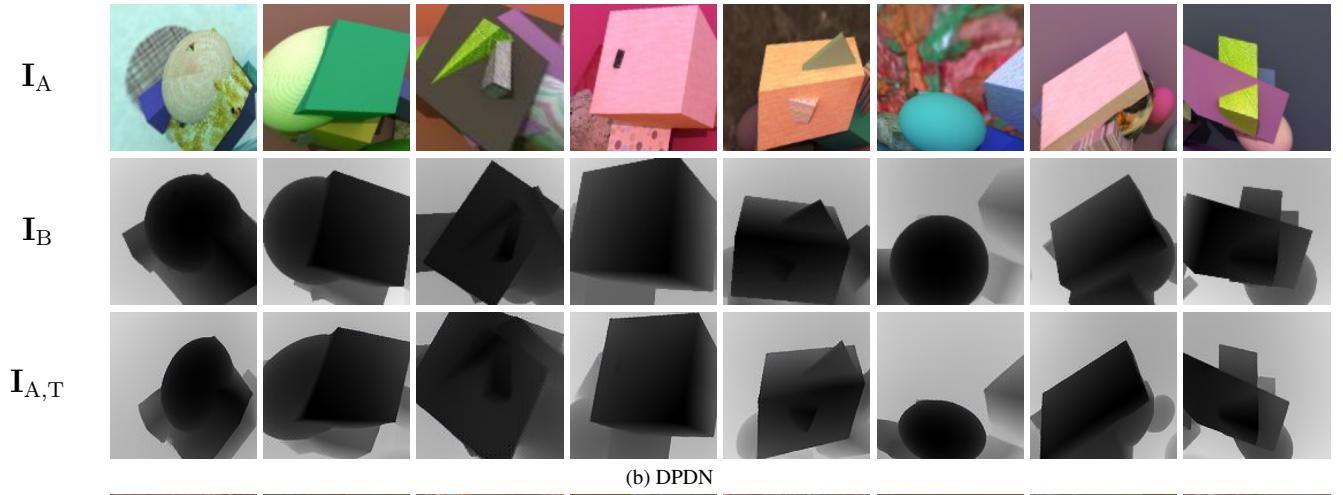
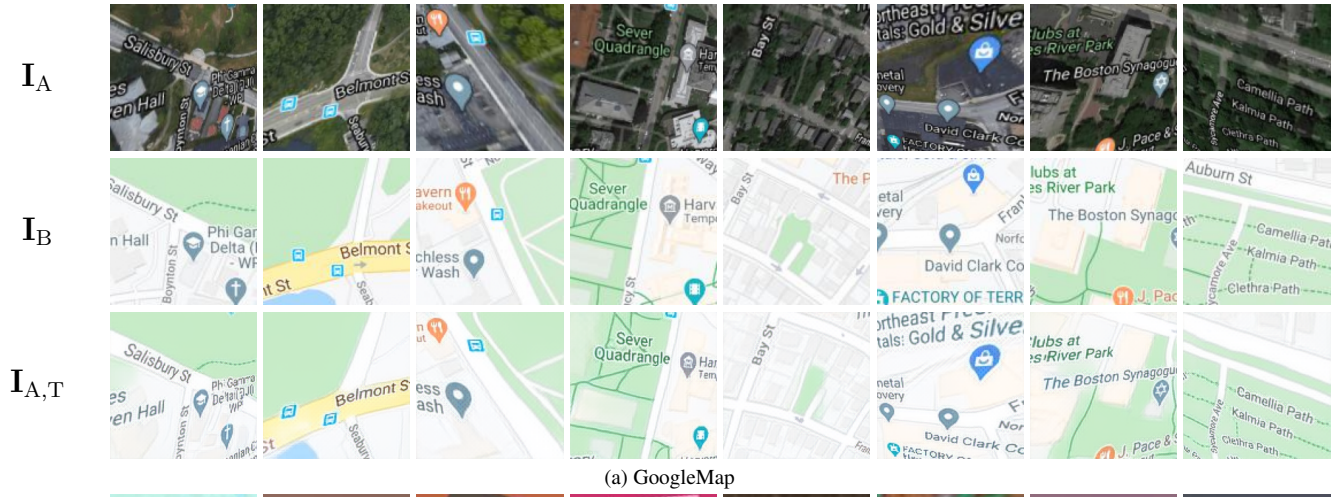
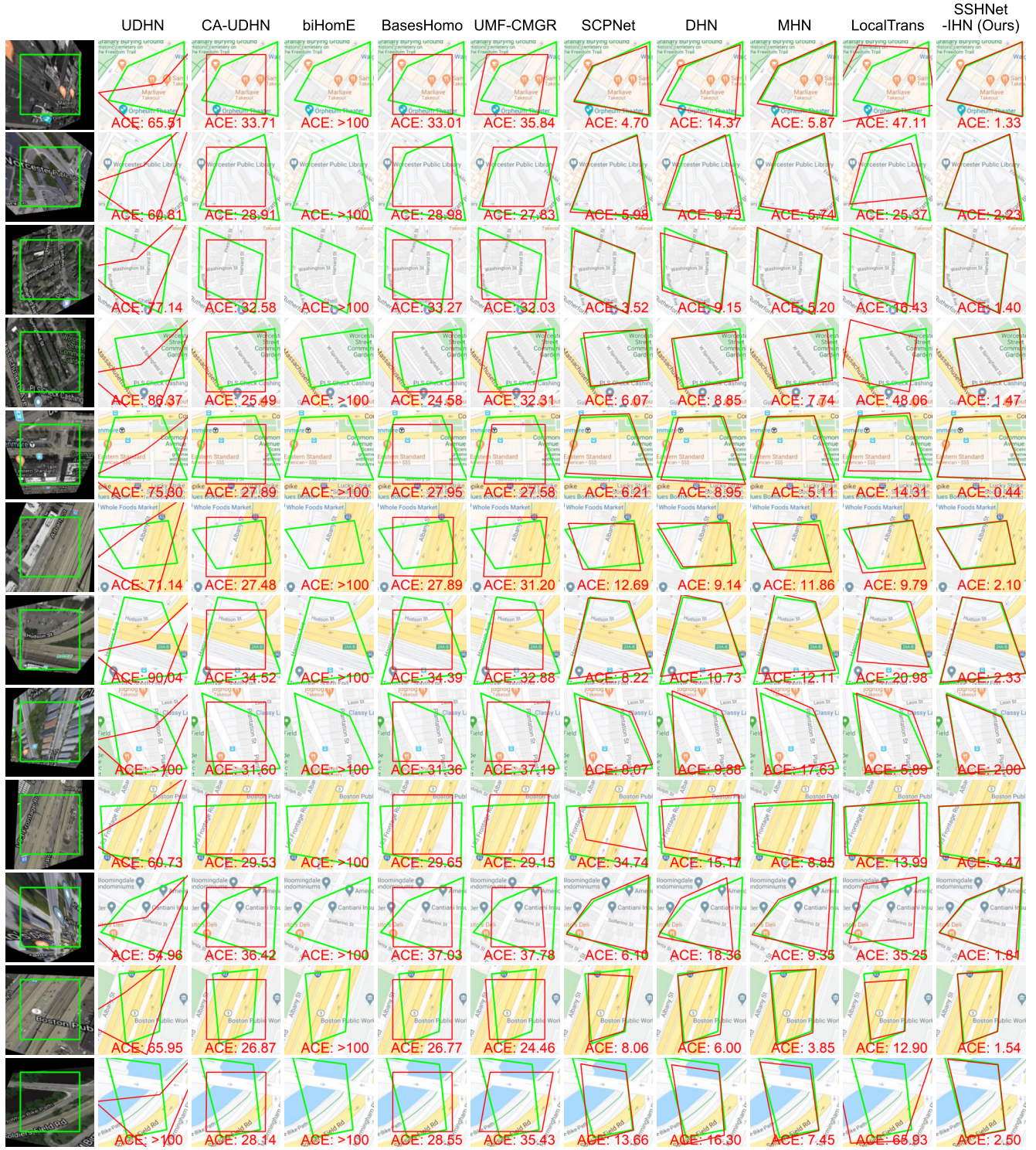


Figure 3. Visualization of modality tranfer results on GoogleMap, DPDN, and OPT-SAR datasets.

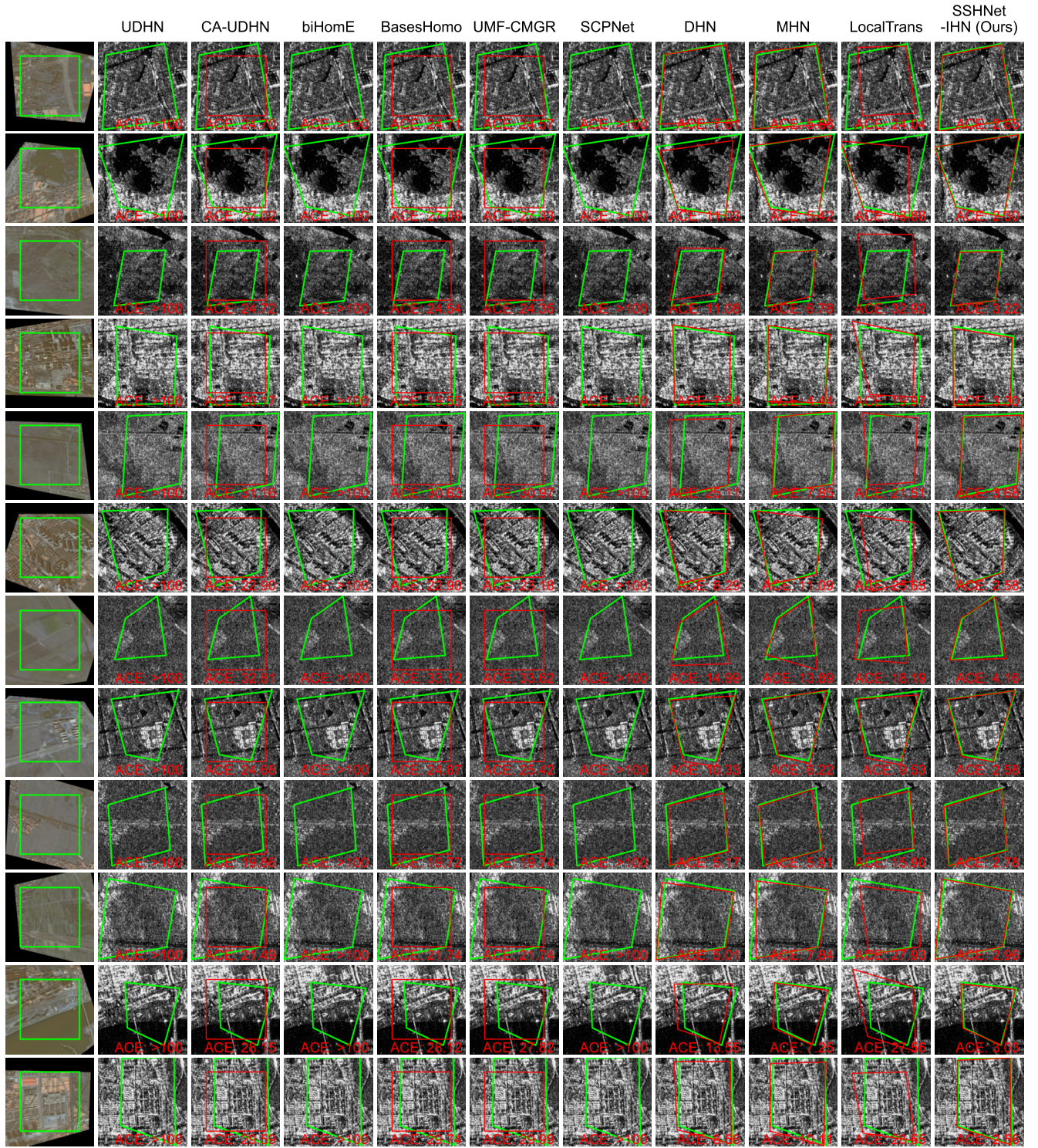
















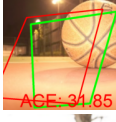









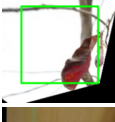





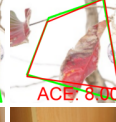



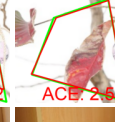











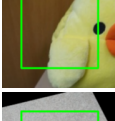



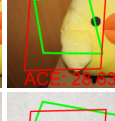
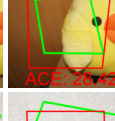


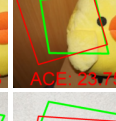

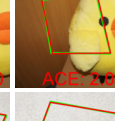
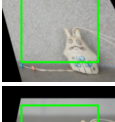


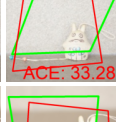
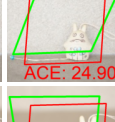
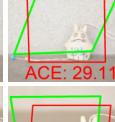


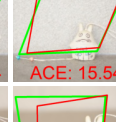
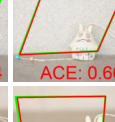
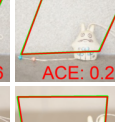
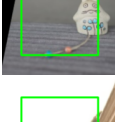

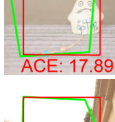
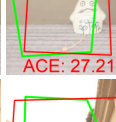
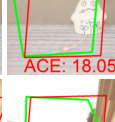
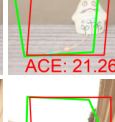
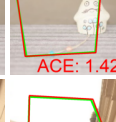
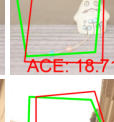
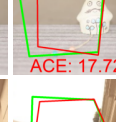
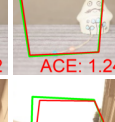
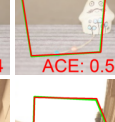
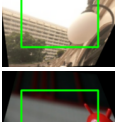









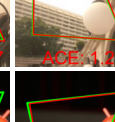
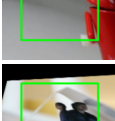









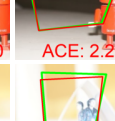
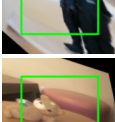
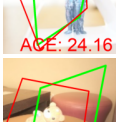









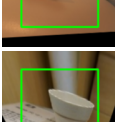










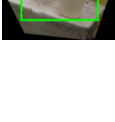
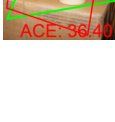
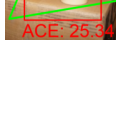
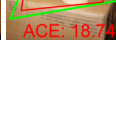
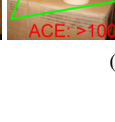


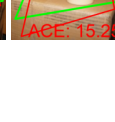
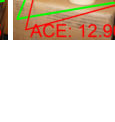
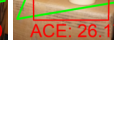

(a) GoogleMap

	UDHN	CA-UDHN	biHomE	BasesHomo	UMF-CMGR	SCPNet	DHN	MHN	LocalTrans	SSHNet -IHN (Ours)
	 ACE: >100	 ACE: 33.45	 ACE: >100	 ACE: 36.96	 ACE: 27.11	 ACE: >100	 ACE: 6.43	 ACE: 3.35	 ACE: 12.40	 ACE: 1.91
	 ACE: >100	 ACE: 29.38	 ACE: >100	 ACE: 31.64	 ACE: 28.36	 ACE: >100	 ACE: 7.89	 ACE: 3.45	 ACE: 2.36	 ACE: 1.44
	 ACE: >100	 ACE: 20.78	 ACE: >100	 ACE: 23.44	 ACE: 28.42	 ACE: >100	 ACE: 7.38	 ACE: 5.50	 ACE: 2.97	 ACE: 1.60
	 ACE: >100	 ACE: 32.15	 ACE: >100	 ACE: 37.44	 ACE: 38.35	 ACE: >100	 ACE: 5.27	 ACE: 6.16	 ACE: 12.82	 ACE: 1.37
	 ACE: >100	 ACE: 22.29	 ACE: >100	 ACE: 22.25	 ACE: 14.70	 ACE: >100	 ACE: 5.84	 ACE: 6.27	 ACE: 2.88	 ACE: 0.88
	 ACE: >100	 ACE: 23.32	 ACE: >100	 ACE: 21.48	 ACE: 26.38	 ACE: >100	 ACE: 5.75	 ACE: 5.42	 ACE: 3.24	 ACE: 1.19
	 ACE: >100	 ACE: >100	 ACE: >100	 ACE: 23.66	 ACE: 26.44	 ACE: >100	 ACE: 8.35	 ACE: 7.90	 ACE: 5.84	 ACE: 1.26
	 ACE: >100	 ACE: >100	 ACE: >100	 ACE: 27.67	 ACE: 32.64	 ACE: >100	 ACE: 13.67	 ACE: 7.55	 ACE: 22.08	 ACE: 4.82
	 ACE: >100	 ACE: 22.10	 ACE: >100	 ACE: 17.26	 ACE: 16.97	 ACE: >100	 ACE: 2.88	 ACE: 4.56	 ACE: 4.05	 ACE: 1.68
	 ACE: >100	 ACE: 25.92	 ACE: >100	 ACE: 25.99	 ACE: 36.92	 ACE: >100	 ACE: 6.85	 ACE: 3.37	 ACE: 4.79	 ACE: 2.03
	 ACE: >100	 ACE: 23.46	 ACE: >100	 ACE: 24.61	 ACE: 27.03	 ACE: >100	 ACE: 5.86	 ACE: 3.94	 ACE: 2.53	 ACE: 1.16
	 ACE: >100	 ACE: 28.35	 ACE: >100	 ACE: 17.47	 ACE: 28.43	 ACE: >100	 ACE: 4.88	 ACE: 6.03	 ACE: 15.81	 ACE: 1.01

(b) DPDN

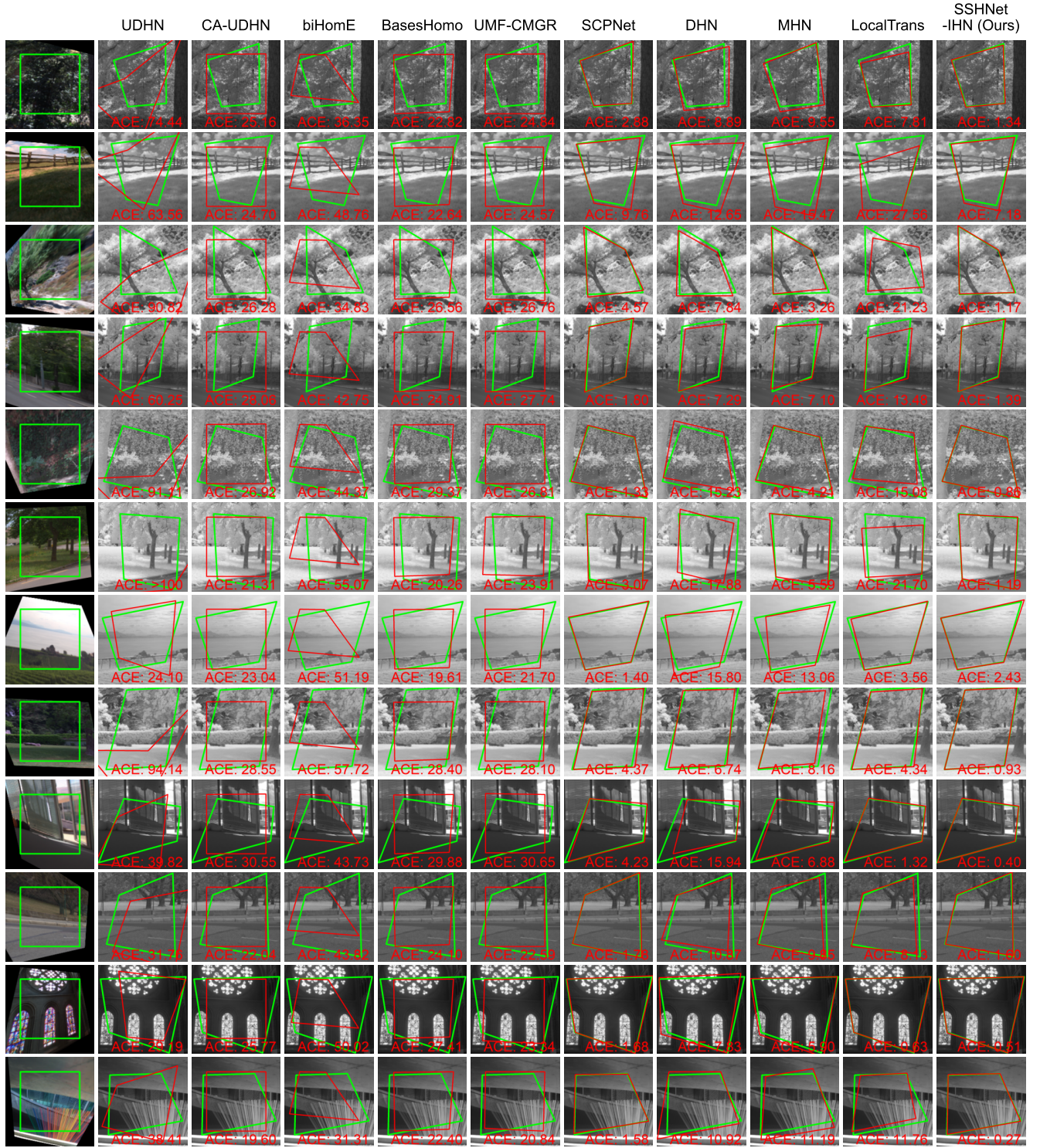


(c) OPT-SAR

	UDHN	CA-UDHN	biHomE	BasesHomo	UMF-CMGR	SCPNet	DHN	MHN	LocalTrans	SSHNet -IHN (Ours)
	 ACE: 38.53	 ACE: 24.34	 ACE: 24.30	 ACE: 24.38	 ACE: 20.27	 ACE: 1.96	 ACE: 7.78	 ACE: 9.85	 ACE: 1.96	 ACE: 0.94
	 ACE: 31.85	 ACE: 28.77	 ACE: 47.57	 ACE: 24.64	 ACE: 34.21	 ACE: 5.92	 ACE: 23.38	 ACE: 17.34	 ACE: 6.43	 ACE: 2.96
	 ACE: 24.38	 ACE: 26.24	 ACE: 35.98	 ACE: 26.63	 ACE: 31.75	 ACE: 8.00	 ACE: 13.92	 ACE: 15.08	 ACE: 8.42	 ACE: 2.52
	 ACE: 41.61	 ACE: 18.05	 ACE: 31.85	 ACE: 22.91	 ACE: 18.98	 ACE: 2.60	 ACE: 20.53	 ACE: 17.64	 ACE: 13.61	 ACE: 1.20
	 ACE: 30.83	 ACE: 35.41	 ACE: 30.64	 ACE: 30.43	 ACE: 30.43	 ACE: 1.61	 ACE: 3.63	 ACE: 3.13	 ACE: 3.32	 ACE: 1.08
	 ACE: 47.26	 ACE: 27.78	 ACE: 33.28	 ACE: 24.90	 ACE: 29.11	 ACE: 1.44	 ACE: 20.54	 ACE: 15.54	 ACE: 0.66	 ACE: 0.29
	 ACE: 23.25	 ACE: 17.89	 ACE: 27.21	 ACE: 18.05	 ACE: 21.26	 ACE: 1.42	 ACE: 18.71	 ACE: 17.72	 ACE: 1.24	 ACE: 0.59
	 ACE: 72.89	 ACE: 47.41	 ACE: 35.02	 ACE: 23.41	 ACE: 20.14	 ACE: 2.48	 ACE: 11.87	 ACE: 6.42	 ACE: 5.37	 ACE: 1.48
	 ACE: 39.82	 ACE: 23.57	 ACE: 32.66	 ACE: >100	 ACE: 24.11	 ACE: 9.21	 ACE: 7.35	 ACE: 12.66	 ACE: 23.40	 ACE: 2.29
	 ACE: 24.16	 ACE: 28.20	 ACE: 31.02	 ACE: >100	 ACE: 31.45	 ACE: 8.52	 ACE: 20.86	 ACE: 16.01	 ACE: 27.15	 ACE: 8.22
	 ACE: 52.05	 ACE: 26.26	 ACE: 38.09	 ACE: >100	 ACE: 27.73	 ACE: 7.88	 ACE: 10.63	 ACE: 5.08	 ACE: 24.59	 ACE: 1.12
	 ACE: 36.40	 ACE: 25.34	 ACE: 18.74	 ACE: >100	 ACE: 26.28	 ACE: 4.91	 ACE: 15.25	 ACE: 12.90	 ACE: 26.17	 ACE: 1.27

(d) Flash/no-flash





(e) RGB/NIR

Figure 4. Qualitative results of homography estimation on GoogleMap, DPDN, OPT-SAR, Flash/no-flash, and RGB/NIR datasets. **Green** polygons denote the ground-truth homography deformation from the deformed source image  $I_A$  to the target image  $I_B$ . **Red** polygons denote the estimated homography deformation using different approaches.