Recurrent Homography Estimation Using Homography-Guided Image Warping and Focus Transformer

Si-Yuan Cao[1,2], Runmin Zhang[2], Lun Luo[2]*, Beinan Yu[2], Zehua Sheng[2], Junwei Li[2], Hui-Liang Shen[2]
[1]Ningbo Innovation Center, Zhejiang University, [2]College of Information Science and Electronic Engineering, Zhejiang University
karlcao@hotmail.com,{runmin_zhang,luolun,yubeinan,shengzehua,lijunwei7788,shenhl}@zju.edu.cn

1. Network Details

We illustrate the network details of our RHWF in this section. The detailed architecture of correlation pooling operation, the backbone and homography aggregator in FocusFormer, the homography parameterization using translation of the 4 corner points, and the homography coordinate projection for the homography-guided image warping are demonstrated.

1.1. Correlation Pooling Operation

As illustrated in Fig. 1, before being sent into the homography aggregator, the correlation volume is separately processed by average-pooling and center sampling in the last 2 dimensions to produce 2 feature maps of the size $H \times W \times R_C \times R_C$ and $H \times W \times (R_C+1) \times (R_C+1)$. The 2 feature maps are then reshaped into $H \times W \times R_C^2$ and $H \times W \times (R_C+1)^2$ and concatenated in the channel dimension and sent into the homography aggregator. This operation keeps the perceptual range while saving the network parameters by reducing the input channel of the feature maps by nearly a half.

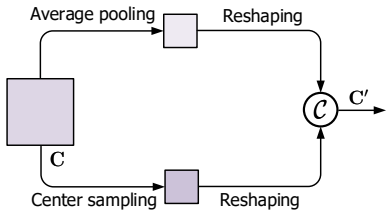

Figure 1. Demonstration of the correlation pooling operation.

1.2. Detailed Architecture of Backbone and Homography Aggregator

The detailed architecture of the backbone and homography aggregator in FocusFormer is shown in Fig. 2. As illustrated in Fig. 2a, the input image is first processed by a 3×3 convolution layer with the depth of 32 and 1 instance normalization+ReLU layer. 2 blocks consisting of 2 residual layers are then used to produce the feature branches for 1-scale RHWF and the additional scale in 2-scale RHWF. The output feature maps of different scales, namely $\mathbf{F}_1$ and $\mathbf{F}_2$, are separately reprojected by two 1×1 convolution layers to achieve a feature channel of 96, which is much fewer than the channel of 256 in IHN [3]. The backbone in our RHWF discards the max-pooling operation that is usually used in the previous works [3, 9]. Instead, RHWF uses the residual layer of stride 2 to conduct the downsampling, which saves the computation costs. We note that the backbone for

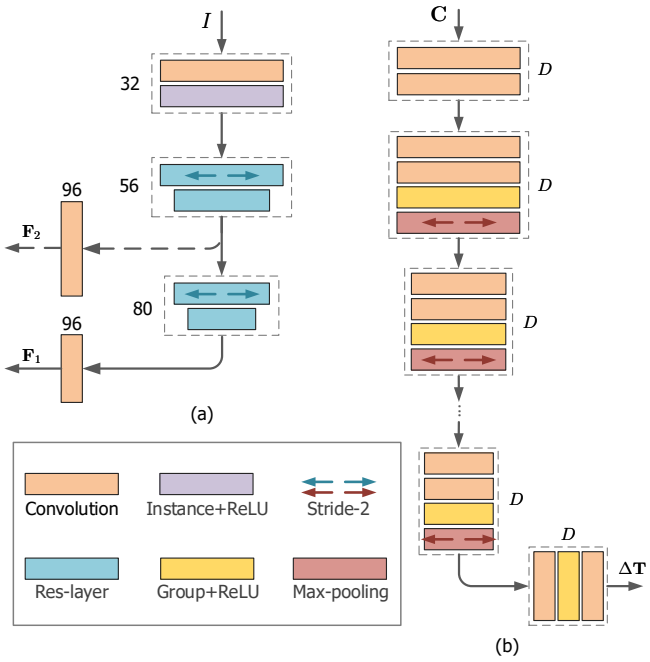

Figure 2. The detailed architecture of the (a) backbone and (b) homography aggregator in FocusFormer. In (a), the solid lines denote the branch for extracting the feature map of 1-scale RHWF, and the dashed lines of the additional scale in 2-scale RHWF. The numbers denote the depth of convolution layers. In (b), D denotes the depth of convolution layers, which is set to $D = 80$ for 1-scale RHWF and $D = 64$ for the extra scale in 2-scale RHWF.

*Corresponding author.

the input image pair I_a and I_b share the same weights, and the feature map of the homography-guided warped image is extracted repeatedly through the recurrence, which can be expressed as

$$\mathbf{F}_b^n = \phi(\mathcal{W}(I_b; \hat{\mathbf{H}}^n)). \tag{1}$$

We illustrate the architecture of the homography aggregator in FocusFormer in Fig. 2b. The correlation is first processed by 2 pure convolution layers and then processed successively by the basic blocks composing of 2 convolution layers, 1 group normalization+ReLU layer, and 1 stride 2 max-pooling layer until the spatial size of the feature map becomes 2×2. The first 2 convolution layers convert the correlation into the latent space to facilitate the estimation. Finally, the feature map is projected by the final block without the max-pooling layer to produce the residual translation prediction $\Delta\mathbf{T}$. We note that the whole FocusFormer works recurrently with tied weights, and hence the homography aggregator works in a similar manner that doesn't require extra parameters during recurrence.

1.3. Homography Parameterization

As in the previous works [3, 5, 6, 9], we parameterize the homography matrix using the translation of the 4 corner points of an image, namely Eq. (6) in the main text. Let's first go over Eq. (6) in the main text as

$$\mathbf{A}^n \hat{\mathbf{h}}^n = \mathbf{b}^n, \tag{2}$$

where $\mathbf{b}^n$ is the coordinate of the projected 4 corner points, and $\mathbf{A}^n$ is composed of the projected 4 corner points and the original 4 corner points. We denote the original 4 corner points of an image as (u_1, v_1), (u_2, v_2), (u_3, v_3), (u_4, v_4), and the corresponding projected ones as (u'_1, v'_1), (u'_2, v'_2), (u'_3, v'_3), (u'_4, v'_4). The two sets of points are related by the predicted $\hat{\mathbf{T}}^n$ as

$$\begin{aligned}
u'^n_1 &= u_1 + \hat{\mathbf{T}}^n(0,0,0) \\
v'^n_1 &= v_1 + \hat{\mathbf{T}}^n(1,0,0) \\
u'^n_2 &= u_2 + \hat{\mathbf{T}}^n(0,0,1) \\
v'^n_2 &= v_2 + \hat{\mathbf{T}}^n(1,0,1) \\
u'^n_3 &= u_3 + \hat{\mathbf{T}}^n(0,1,0) \\
v'^n_3 &= v_3 + \hat{\mathbf{T}}^n(1,1,0) \\
u'^n_4 &= u_4 + \hat{\mathbf{T}}^n(0,1,1) \\
v'^n_4 &= v_4 + \hat{\mathbf{T}}^n(1,1,1).
\end{aligned} \tag{3}$$

And then we construct $\mathbf{A}^n$ as

$$\mathbf{A}^n = \begin{bmatrix}
u_1 & v_1 & 1 & 0 & 0 & 0 & -u_1 u'^n_1 & -v_1 u'^n_1 \\
0 & 0 & 0 & u_1 & v_1 & 1 & -u_1 v'^n_1 & -v_1 v'^n_1 \\
u_2 & v_2 & 1 & 0 & 0 & 0 & -u_2 u'^n_2 & -v_2 u'^n_2 \\
0 & 0 & 0 & u_2 & v_2 & 1 & -u_2 v'^n_2 & -v_2 v'^n_2 \\
u_3 & v_3 & 1 & 0 & 0 & 0 & -u_3 u'^n_3 & -v_3 u'^n_3 \\
0 & 0 & 0 & u_3 & v_3 & 1 & -u_3 v'^n_3 & -v_3 v'^n_3 \\
u_4 & v_4 & 1 & 0 & 0 & 0 & -u_4 u'^n_4 & -v_4 u'^n_4 \\
0 & 0 & 0 & u_4 & v_4 & 1 & -u_4 v'^n_4 & -v_4 v'^n_4
\end{bmatrix}, \tag{4}$$

and $\mathbf{b}^n$ as

$$\mathbf{b}^n = \begin{bmatrix} u'^n_1 & v'^n_1 & u'^n_2 & v'^n_2 & u'^n_3 & v'^n_3 & u'^n_4 & v'^n_4 \end{bmatrix}^\mathsf{T}. \tag{5}$$

Finally, the vectorized homography can be expressed as

$$\hat{\mathbf{h}}^n = \begin{bmatrix} \hat{\mathbf{H}}^n_{11} & \hat{\mathbf{H}}^n_{12} & \hat{\mathbf{H}}^n_{13} & \hat{\mathbf{H}}^n_{21} & \hat{\mathbf{H}}^n_{22} & \hat{\mathbf{H}}^n_{23} & \hat{\mathbf{H}}^n_{31} & \hat{\mathbf{H}}^n_{32} \end{bmatrix}^\mathsf{T}, \tag{6}$$

which can be computed by Eq. (6) in the main text.

1.4. Homography Coordinate Projection

Once we obtain the homography matrix $\hat{\mathbf{H}}^n$, the original coordinate of an image $\mathbf{x} = (u, v)$ can be projected by

$$\begin{bmatrix} u'^n \\ v'^n \\ 1 \end{bmatrix} \sim \begin{bmatrix} \hat{\mathbf{H}}^n_{11} & \hat{\mathbf{H}}^n_{12} & \hat{\mathbf{H}}^n_{13} \\ \hat{\mathbf{H}}^n_{21} & \hat{\mathbf{H}}^n_{22} & \hat{\mathbf{H}}^n_{23} \\ \hat{\mathbf{H}}^n_{31} & \hat{\mathbf{H}}^n_{32} & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}. \tag{7}$$

The projection is conducted on each coordinate position of an image, which enables the pixel-wised homography-guided image warping, namely $\mathcal{W}(I_b, \hat{\mathbf{H}}^n)$. We note that, in practice, the computed homography needs to be scaled to perform the image warping. Taking the homography calculated on the 32×32 feature map as an example, the scale should be multiplied by a factor of 4.

2. Dataset Details

We evaluate our RHWF on the datasets including MSCOCO [7], 4×/8× cross-resolution MSCOCO [9], and GoogleEarth/GoogleMap cross-modal [10] datasets. For a fair comparison, all the methods included for evaluation are trained and tested by the same training and test splits on each dataset. It is worth noting that in the previous works CLKN and DLKFM [4, 10], one of the input images is of a larger size 192×192. Previous work LocalTrans [9] impaints the boundary of the warped image by warping the image of a larger size 192×192 and re-cropping it. On the other side, our RHWF only needs the input images of size 128×128, and the warping is also conducted on the image of size 128×128, which means our RHWF achieves better accuracy with less information. Fig. 2 shows the input image pairs of each dataset. In the following, we will illustrate each dataset in detail.

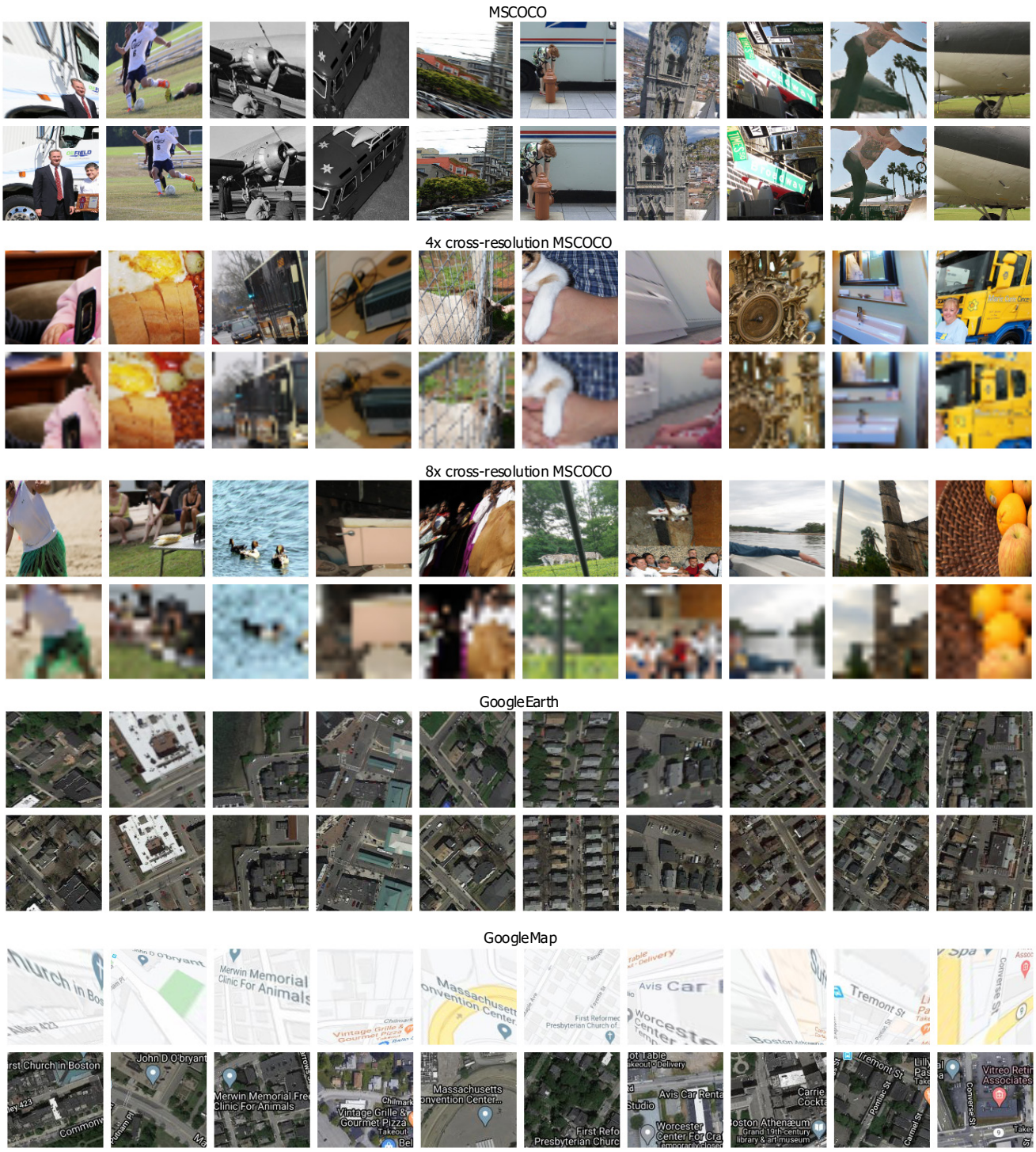

Figure 2. Exemplary image pairs related by homography of different datasets.

MSCOCO [7] is a large-scale image dataset of real-world RGB images, which is first used in [5] for deep homography estimation. The input image is uniformly resized into 320×240 and then perturbed to form the image under the homography deformation. The image pair of size 128×128 is then produced by cropping the 320×240 images at the same position. The perturbation range is set to $[-32, 32]$ as in [3–6, 9, 10].

Cross-Resolution MSCOCO is first employed in [9]. The homography estimation of this kind of image data facilitates multiscale gigapixel photography [2]. The data generation process of the cross-resolution MSCOCO is basically similar to MSCOCO, except for one of the produced 128×128 images is first downsampled by the factor of 4 or 8 and then upsampled to 128×128 to produce the image pairs of resolution gaps.

GoogleEarth contains high-resolution satellite images captured on different dates. The images captured for the same place in different seasons are cropped into 192×192 image pairs, which enables the 128×128 images to have a perturbation of $[-32, 32]$. The satellite images saved on 04/2018 and 06/2019 about the Great Boston area are used to build the dataset, which brings the image pair modality gaps by introducing the temporal change. It is worth noting that the training split of GoogleEarth fixes the transformation of image pairs, which limits the homography deformation augmentation in the training stage. Consequently, our RHWF comes into overfitting after 30000 training iterations, and hence we use an early stopping strategy by taking the model at the 30000th iteration as the final model for GoogleEarth. For other methods, we employ their best reported results on GoogleEarth for a fair comparison.

GoogleMap contains multimodal images provided by the Google Static Map API. Two corresponding images belong to static google map and satellite map are cropped to form the multimodal image pair. The cropped image size is 192×192 as in GoogleEarth. It is observed that severe modality gap exists in the image pairs.

3. More Experimental Results

We further illustrate more experimental results of our RHWF, including the recurrent ACE comparison with the previous SOTA work IHN [3], the attention map of the attention-focusing mechanism at each iteration, and the homography estimation results.

3.1. Recurrent ACE Comparison

The homography estimation errors, namely ACE, during the recurrent process can further reveal the character of our proposed RHWF. We compare the ACE during the recurrent process of the previous SOTA work IHN [3] and our RHWF in Fig. 3. It is interesting that our RHWF might produce the estimation with less accuracy at the former iterations, while it can surpass IHN after few iterations. One possible explanation is that homography-guided image warping progressively reduces the feature inconsistency and the attention-focusing mechanism aggregates the intra/inter correspondence information in a gradually focusing manner through the iteration.

3.2. Attention Map of the Attention-Focusing Mechanism

In Fig. 4, 5, and 6, we illustrate the attention maps of the attention-focusing mechanism at each iteration on cross-resolution MSCOCO. It is observed that as the resolution gap grows, the global attention map becomes more ambiguous. Fortunately, our proposed homography-guided image warping can correct the deformation between the 2 images and the attention-focusing mechanism gradually shrinks the attention range, which can clarify the attention.

3.3. Homography Estimation Results

We demonstrate the homography estimation results in Fig. 7 and 8. It is observed that under the severe homography deformation and resolution/modality gaps, our RHWF provides stable and accurate homography estimation results. It is interesting in Fig. 7 that the scenes at the 2nd and 3rd row of the 4x cross-resolution MSCOCO are extremely lack of texture, but our RHWF is still able to estimate the homography accurately, which further reveals the effectiveness of the homography-guided image warping and FocusFormer.

References

[1] Daniel Barath, Jiri Matas, and Jana Noskova. MAGSAC: marginalizing sample consensus. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10197–10205, 2019. 10

[2] David J Brady, Michael E Gehm, Ronald A Stack, Daniel L Marks, David S Kittle, Dathon R Golish, EM Vera, and Steven D Feller. Multiscale gigapixel photography. *Nature*, 486(7403):386–389, 2012. 4

[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, 2, 3, 4, 5, 10

[4] Che-Han Chang, Chun-Nan Chou, and Edward Y Chang. CLKN: Cascaded lucas-kanade networks for image alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2213–2221, 2017. 2, 3, 10

[5] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Deep image homography estimation. *arXiv preprint arXiv:1606.03798*, 2016. 2, 3, 9, 10

[6] 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, 3, 9, 10

[7] 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 *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014. 2, 3

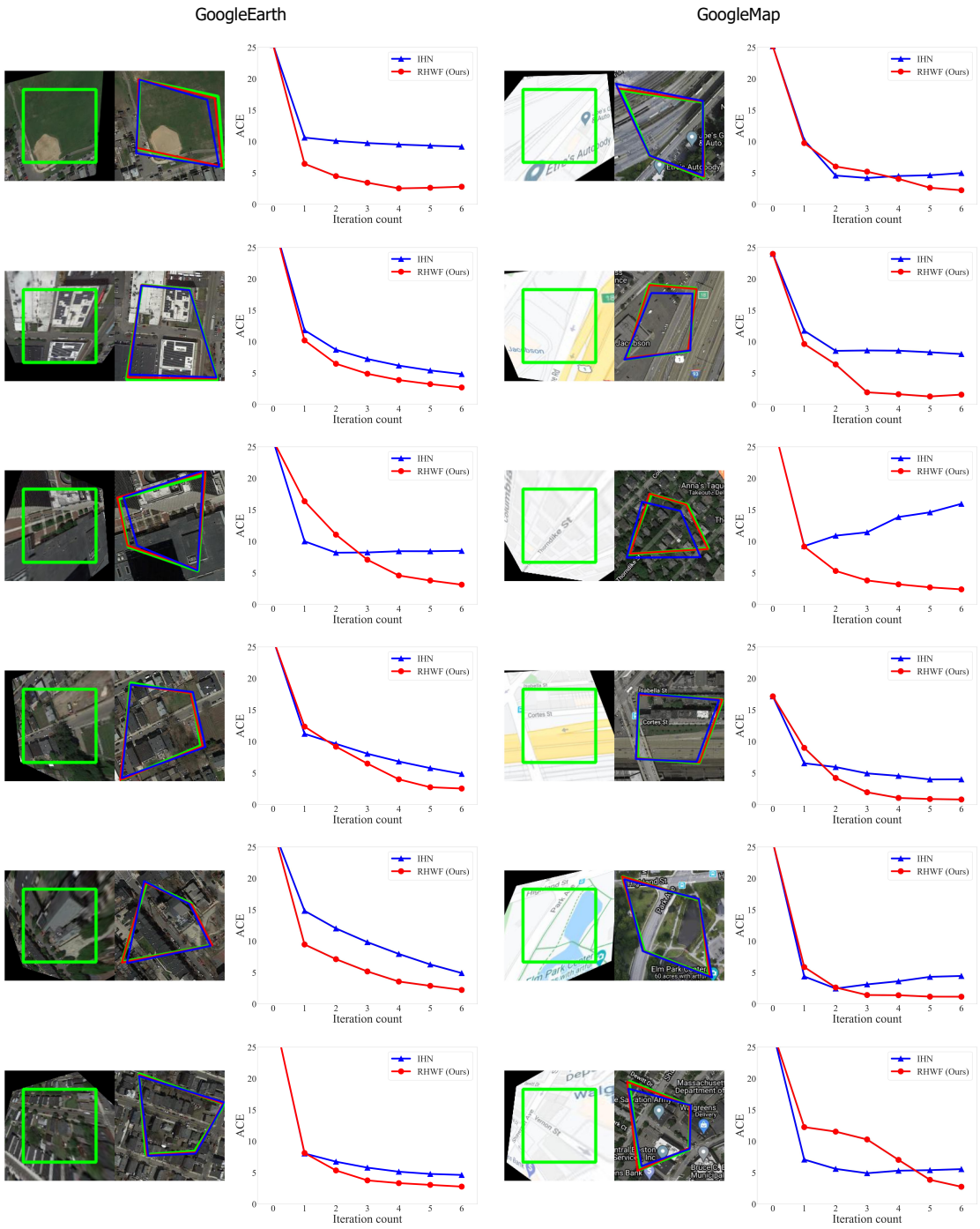

Figure 3. The recurrent ACE comparison with the previous SOTA work IHN [3].

[8] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 9, 10

[9] 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. 1, 2, 3, 4, 9

[10] 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,

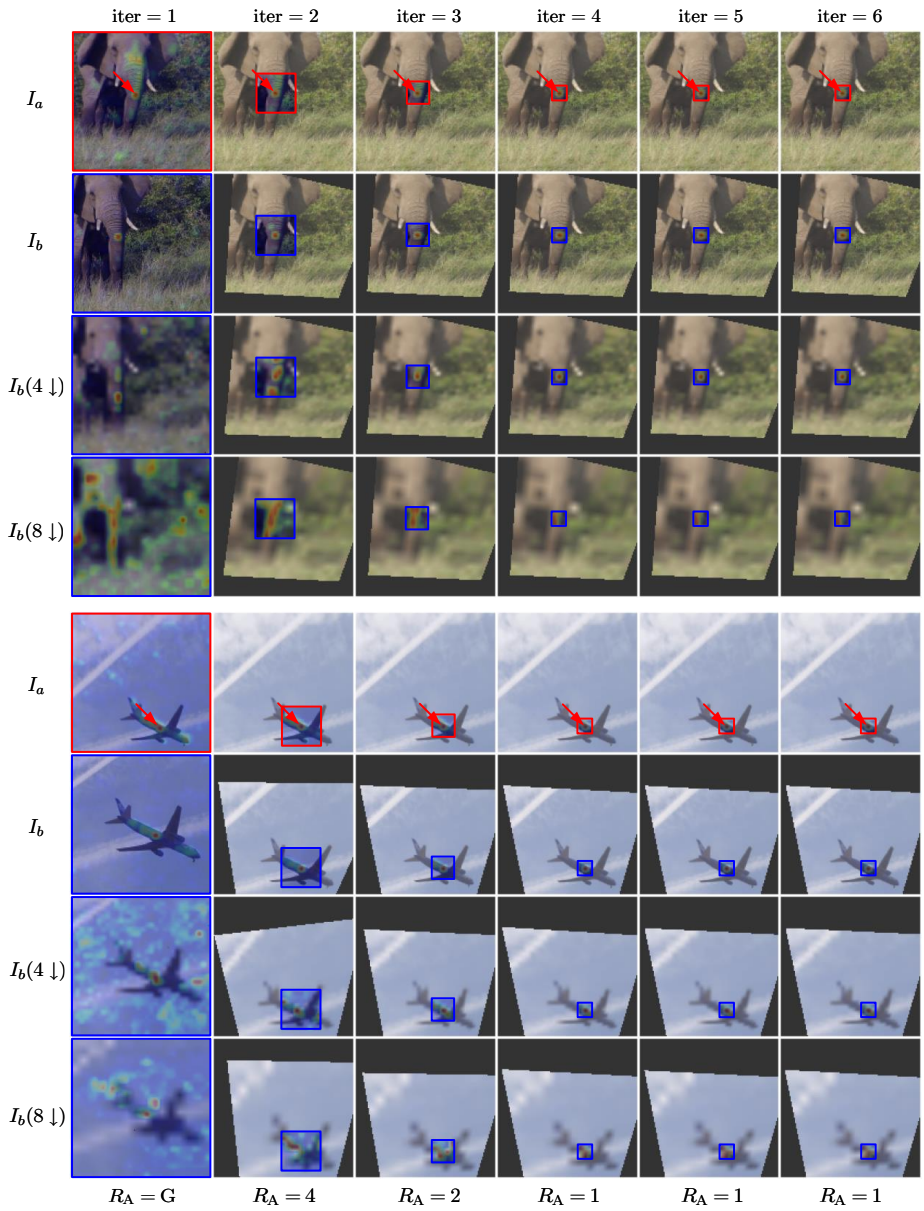

Figure 4. The self- and cross-attention map of RHWF at each iteration. For each scene, The 1st row shows the image I_a of the standard resolution with the self-attention map. The 2nd, 3rd and 4th row shows the image I_b with no (I_b), $4\times$ ($I_b(4\downarrow)$), and $8\times$ ($I_b(8\downarrow)$) downsampling and cross-attention maps. The red arrows denote the query point of attention, and the red and blue boxes separately highlight the self- and cross-attention maps.

2021. 2, 3, 10

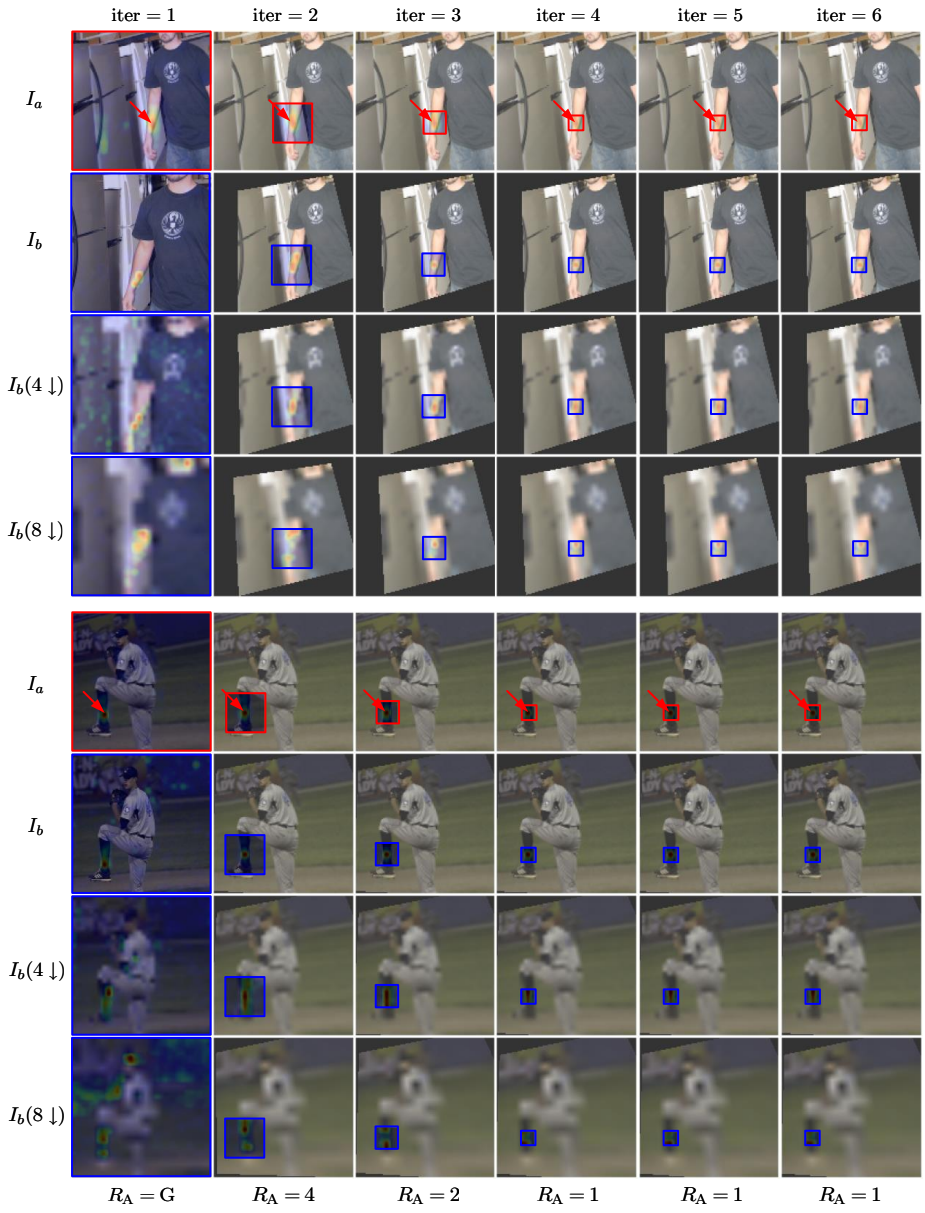

Figure 5. The self- and cross-attention map of RHWF at each iteration. For each scene, The 1st row shows the image I_a of the standard resolution with the self-attention map. The 2nd, 3rd and 4th row shows the image I_b with no (I_b), $4\times$ ($I_b(4\downarrow)$), and $8\times$ ($I_b(8\downarrow)$) downsampling and cross-attention maps. The red arrows denote the query point of attention, and the red and blue boxes separately highlight the self- and cross-attention maps.

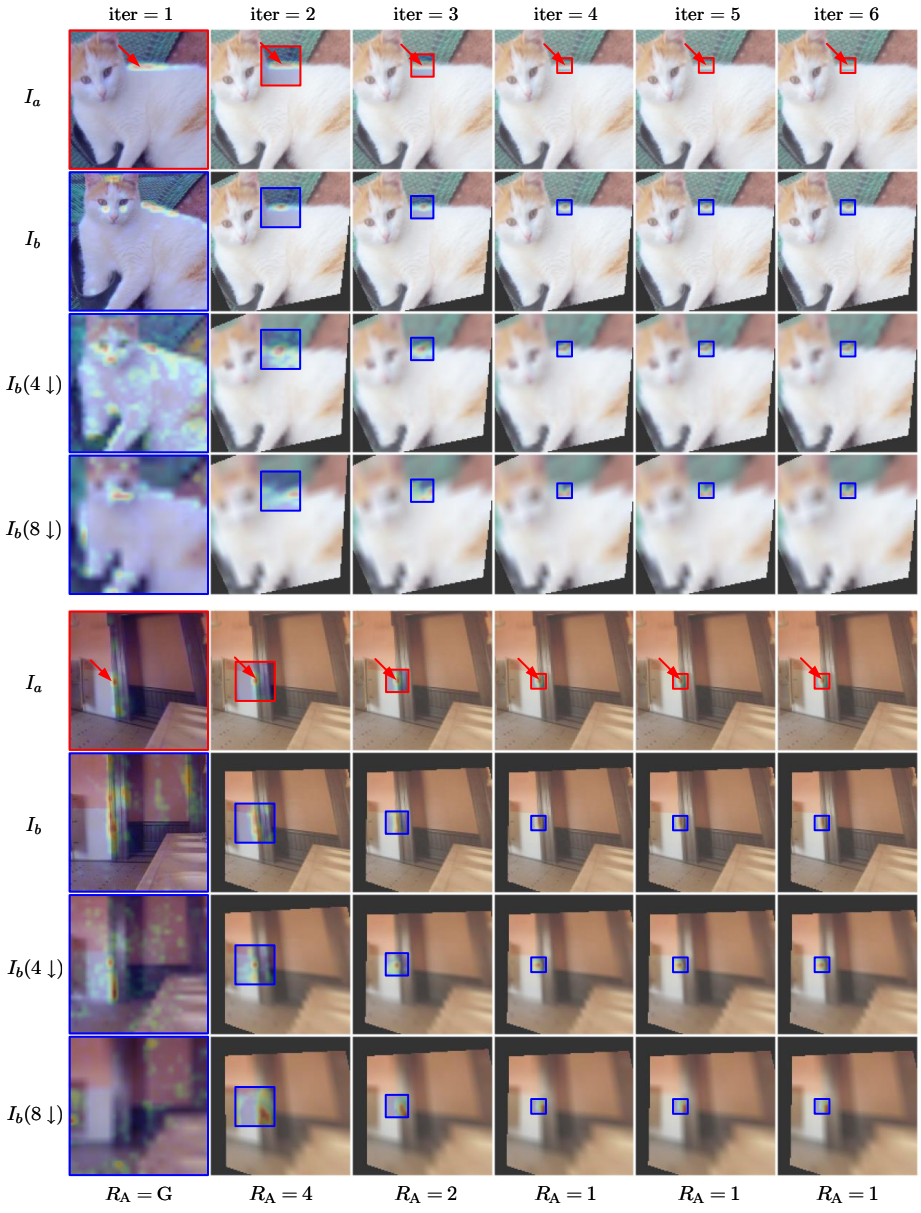

Figure 6. The self- and cross-attention map of RHWF at each iteration. For each scene, The 1st row shows the image I_a of the standard resolution with the self-attention map. The 2nd, 3rd and 4th row shows the image I_b with no (I_b), $4\times$ ($I_b(4\downarrow)$), and $8\times$ ($I_b(8\downarrow)$) downsampling and cross-attention maps. The red arrows denote the query point of attention, and the red and blue boxes separately highlight the self- and cross-attention maps.

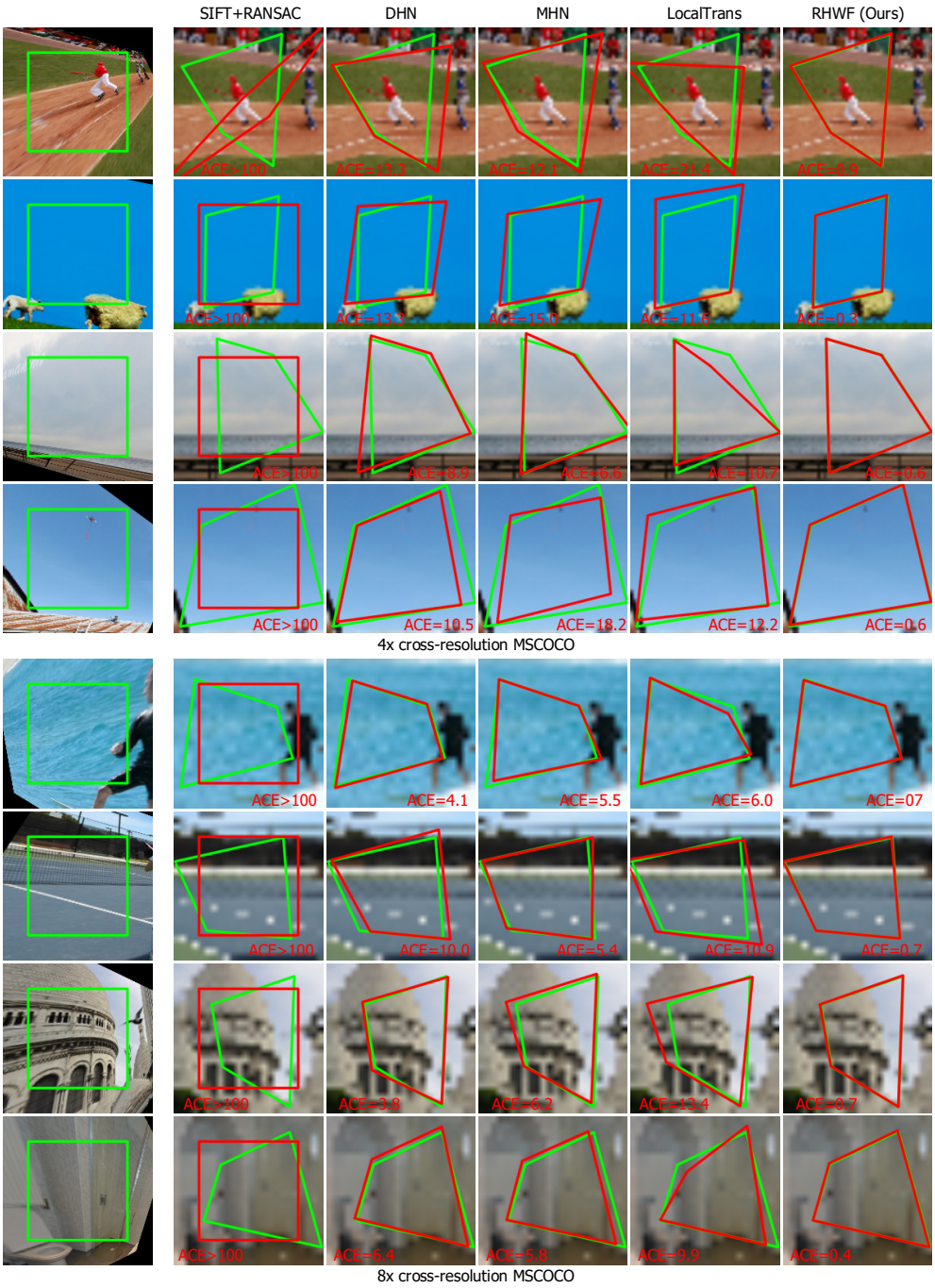

Figure 7. Homography estimation results of methods including SIFT+RANSAC [8], DHN [5], MHN [6], LocalTrans [9], and our RHWF.

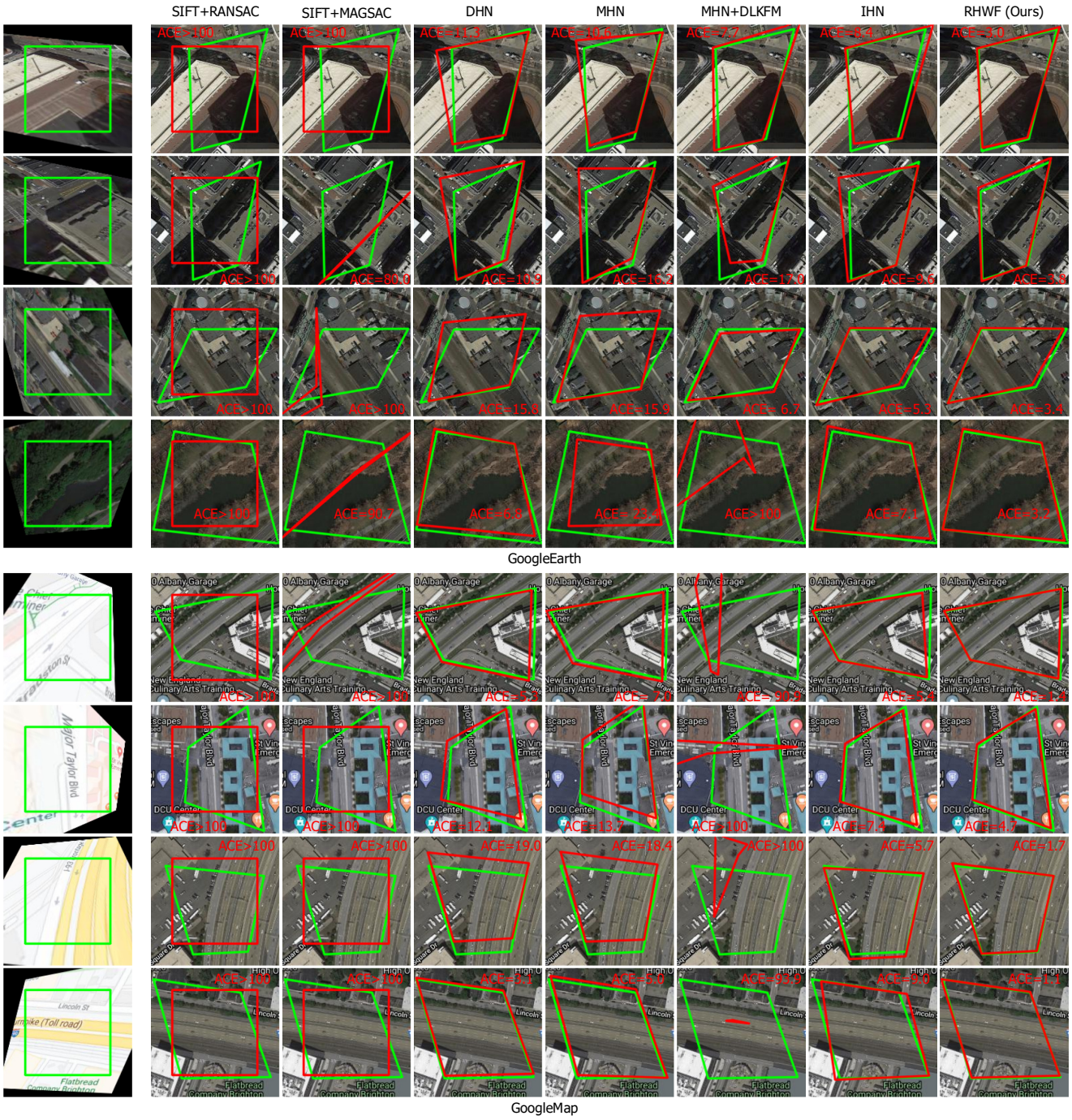

Figure 8. Homography estimation results of methods including IHN [3], SIFT+RANSAC [8], SIFT+MAGSAC [1], CLKN [4], DHN [5], MHN [6], MHN+DLKFM [10], and our RHWF.