MCNet: Rethinking the Core Ingredients for Accurate and Efficient Homography Estimation Supplementary Materials

1. Homography Parameterization

We parameterize the homography matrix $\mathbf{M}$ using the translation $\mathbf{T}$ of the 4 corner points based on the least squares method, which can be expressed as

$$\mathbf{Am} = \mathbf{b}, \tag{1}$$

where $\mathbf{A}$ is composed of the original coordinates of four points on the source image $\mathbf{I}_{\mathrm{S}}$ and the mapped coordinates on the target image $\mathbf{I}_{\mathrm{T}}$, $\mathbf{b}$ represents the mapped coordinates of four points, and $\mathbf{m}$ denotes the vectorized homography matrix. We denote the original coordinate as (u_i, v_i) and the mapped coordinate as (u_i', v_i'), where i ranges from 1 to 4, representing four points. Then the mapped coordinates of four points using the estimated translation $\mathbf{T}$ can be expressed as

$$\begin{aligned}
u'_1 &= u_1 + \mathbf{T}(0,0,0) \\
v'_1 &= v_1 + \mathbf{T}(1,0,0) \\
u'_2 &= u_2 + \mathbf{T}(0,0,1) \\
v'_2 &= v_2 + \mathbf{T}(1,0,1) \\
u'_3 &= u_3 + \mathbf{T}(0,1,0) \\
v'_3 &= v_3 + \mathbf{T}(1,1,0) \\
u'_4 &= u_4 + \mathbf{T}(0,1,1) \\
v'_4 &= v_4 + \mathbf{T}(1,1,1).
\end{aligned} \tag{2}$$

Given the homography matrix $\mathbf{M}$, the relationship between the original four points (u_i, v_i) and the mapped four points (u_i', v_i') can be represented as

$$\begin{aligned}
u_i' &= \frac{\mathbf{M}_{11}u_i + \mathbf{M}_{12}v_i + \mathbf{M}_{13}}{\mathbf{M}_{31}u_i + \mathbf{M}_{32}v_i + 1} \\
v_i' &= \frac{\mathbf{M}_{21}u_i + \mathbf{M}_{22}v_i + \mathbf{M}_{23}}{\mathbf{M}_{31}u_i + \mathbf{M}_{32}v_i + 1}.
\end{aligned} \tag{3}$$

The above equations can be rearranged to be

$$\begin{aligned}
u_i' &= \mathbf{M}_{11}u_i + \mathbf{M}_{12}v_i + \mathbf{M}_{13} - \mathbf{M}_{31}u_iu_i' - \mathbf{M}_{32}v_iu_i' \\
v_i' &= \mathbf{M}_{21}u_i + \mathbf{M}_{22}v_i + \mathbf{M}_{23} - \mathbf{M}_{31}u_iv_i' - \mathbf{M}_{32}v_iv_i'.
\end{aligned} \tag{4}$$

Then we construct the matrix $\mathbf{A}$ as

$$\mathbf{A} = \begin{bmatrix}
u_1 & v_1 & 1 & 0 & 0 & 0 & -u_1u_1' & -v_1u_1' \\
0 & 0 & 0 & u_1 & v_1 & 1 & -u_1v_1' & -v_1v_1' \\
u_2 & v_2 & 1 & 0 & 0 & 0 & -u_2u_2' & -v_2u_2' \\
0 & 0 & 0 & u_2 & v_2 & 1 & -u_2v_2' & -v_2v_2' \\
u_3 & v_3 & 1 & 0 & 0 & 0 & -u_3u_3' & -v_3u_3' \\
0 & 0 & 0 & u_3 & v_3 & 1 & -u_3v_3' & -v_3v_3' \\
u_4 & v_4 & 1 & 0 & 0 & 0 & -u_4u_4' & -v_4u_4' \\
0 & 0 & 0 & u_4 & v_4 & 1 & -u_4v_4' & -v_4v_4'
\end{bmatrix}, \tag{5}$$

and the mapped coordinates $\mathbf{b}$ as

$$\mathbf{b} = \begin{bmatrix} u'_1 & v'_1 & u'_2 & v'_2 & u'_3 & v'_3 & u'_4 & v'_4 \end{bmatrix}^{\mathsf{T}}. \tag{6}$$

Finally, the vectorized homography matrix can be expressed as

$$\mathbf{m} = \begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} & \mathbf{M}_{13} & \mathbf{M}_{21} & \mathbf{M}_{22} & \mathbf{M}_{23} & \mathbf{M}_{31} & \mathbf{M}_{32} \end{bmatrix}^{\mathsf{T}}. \tag{7}$$

The final homography matrix $\mathbf{M}$ is obtained by solving the least squares problem.

2. Dataset Details

We evaluate our MCNet on MSCOCO [7], GoogleEarth [11], GoogleMap [11], and SPID [9] datasets, as shown in Fig. 1. The size of input image pairs is set to be 128×128. In the following, we will explain the details for each dataset.

MSCOCO. MSCOCO is a widely used large-scale image dataset in computer vision tasks, which covers a variety of common scenarios, serving as a fundamental dataset for evaluating homography estimation methods. We process MSCOCO as in [2–6, 10, 11]. The images are first resized to 320×240. Then we crop a 128×128 patch from the resized image as the target image. The resized image is then deformed using the homography transformation produced by the random perturbation of four corner points within the range $[-32, 32]$. The same region of the deformed image is cropped to serve as the source image.

GoogleEarth. GoogleEarth consists of cross-season satellite images collected on 04/2018 and 06/2019 in the Great Boston area. We process GoogleEarth in the same

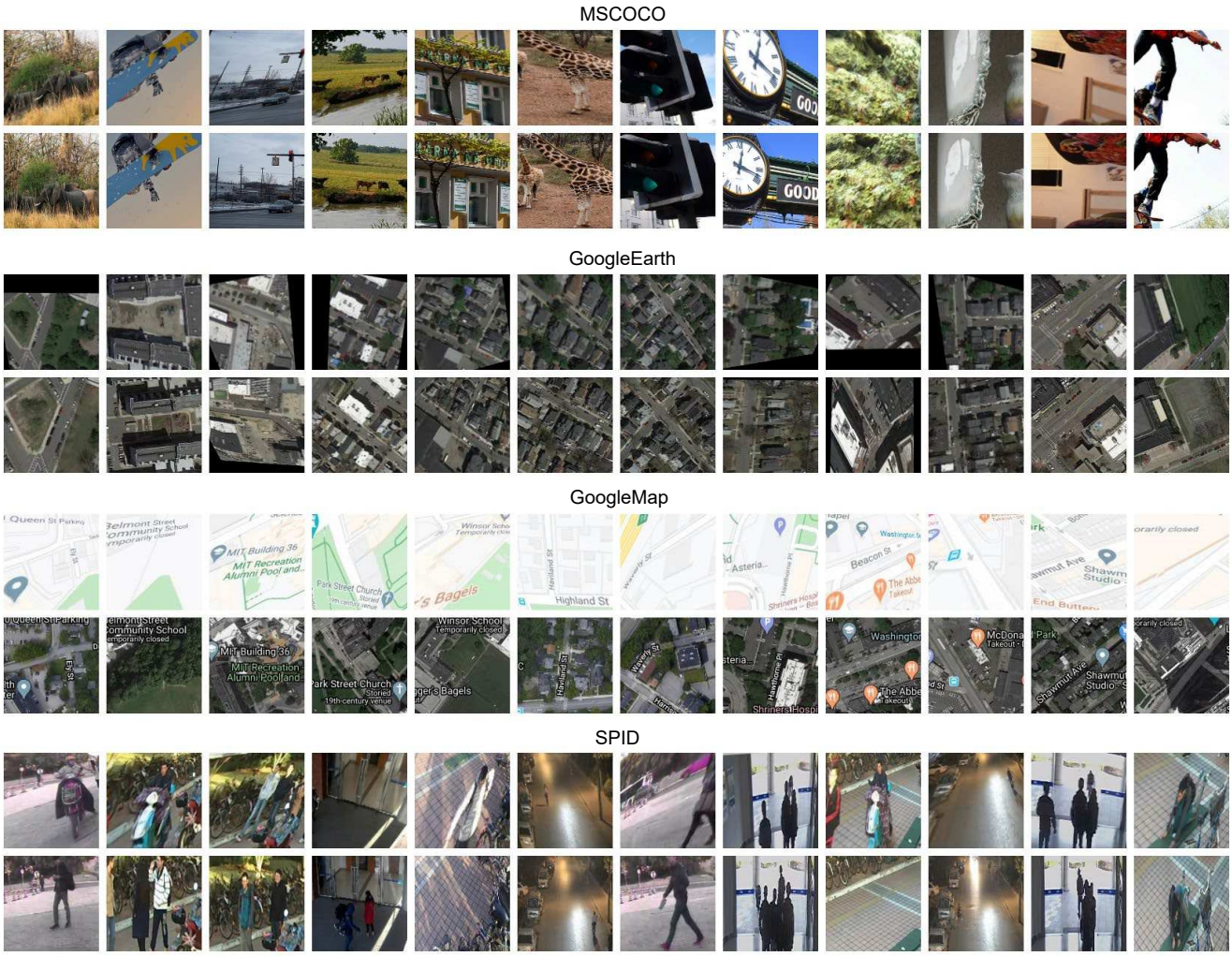

Figure 1. Example image pairs of MSCOCO, Google Earth, Google Map, and SPID datasets. MSCOCO dataset consists of common RGB images. Google Earth and Google Map datasets contain data from cross-modalities. SPID dataset specifically provides surveillance images that include dynamic foreground objects.

way as in [2, 3, 11]. The images are cropped into 192×192 image pairs, which then enables the 128×128 images to have a perturbation of $[-32, 32]$.

GoogleMap. GoogleMap includes satellite images and corresponding map images of the same region. We process GoogleMap in the same way as in [2, 3, 11]. The images are cropped into 192×192 image pairs, which then enables the 128×128 images to have a perturbation of $[-32, 32]$.

SPID. SPID dataset provides surveillance images that include dynamic foreground objects. We process the dataset in the same way as in [2]. The original image pairs are cropped to 220×220, and then produce the 128×128 images to have a perturbation of $[-32, 32]$.

3. More Experimental Results

This section presents more experimental results on each dataset, including the average corner error (ACE) at each iteration, homography estimation result, and correlation at each iteration. To ensure a fair comparison, all algorithms are trained using the same data processing and splitting strategy.

ACE at each Iteration. Our MCNet is compared with the two previous deep iteration-based methods 2-scale RHWF [3] and 2-scale IHN [2] in terms of ACE at each iteration, to better demonstrate the effectiveness of our model. As illustrated in Fig. 2, our MCNet consistently achieves observably more accurate estimation as the iteration continues, while the error reduction of 2-scale RHWF and 2-scale IHN generally becomes inconspicuous as the iteration grows.

Homography estimation result. We further demonstrate the homography estimation results on each dataset for various methods. For GoogleEarth and GoogleMap datasets, we compare our MCNet with 2-scale RHWF [3], 2-scale IHN [2], MHN+DLKFM [11], DHN [5], MHN [6], SIFT+MAGSAC [1], and SIFT+RANSAC [8], as shown in Fig. 3 and Fig. 4. It is observed that our MCNet achieves highly stable and accurate estimation results. For the SPID dataset, we compare our MCNet with IHN [2], UDHN [10], MHN [6], DHN [5], SIFT+MAGSAC [1], and SIFT+RANSAC [8], as shown in Fig. 5, where UDHN is trained in a supervised manner as in [5]. We note that MCNet shows high estimation accuracy despite the existence of dynamic foreground objects.

Correlation at each Iteration. To further demonstrate the superior accuracy achieved by our MCNet, which is attributed to its multiscale correlation searching design, we visualize the correlations obtained in each iteration for correlation-based methods, including our MCNet, the previous state-of-the-art (SOTA) method RHWF [3], and IHN [2]. The obtained correlations have a channel dimension of $(2r+1) \times (2r+1)$, where r represents the searching radius. We then visualize the center correlation, which is in the $\lfloor (2r+1)^2/2 \rfloor$th dimension of the correlation. In the visualization, the darker regions indicate lower correlation activation values, which are expected to be the areas with significant differences. As shown in Fig. 6 and Fig. 7, for the GoogleEarth and GoogleMap datasets, with the continuation of iterations, the correlation quality of RHWF and IHN remains unimproved, while MCNet consistently achieves better correlation quality. For the SPID dataset, the correlations of MCNet reject outliers in the foreground region with better accuracy than IHN, as shown in Fig. 8, which can significantly improve the estimation accuracy.

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. 3, 5, 6, 7

[2] 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, 6, 7, 8, 9, 10

[3] 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, 3, 4, 5, 6, 8, 9

[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.

[5] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Deep image homography estimation. *arXiv preprint arXiv:1606.03798*, 2016. 3, 5, 6, 7

[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. 1, 3, 5, 6, 7

[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. 1

[8] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 3, 5, 6, 7

[9] Dan Wang, Chongyang Zhang, Hao Cheng, Yanfeng Shang, and Lin Mei. SPID: Surveillance pedestrian image dataset and performance evaluation for pedestrian detection. In *Asian Conference on Computer Vision*, pages 463–477. Springer, 2016. 1

[10] 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. 1, 3, 7

[11] 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, 3, 5, 6

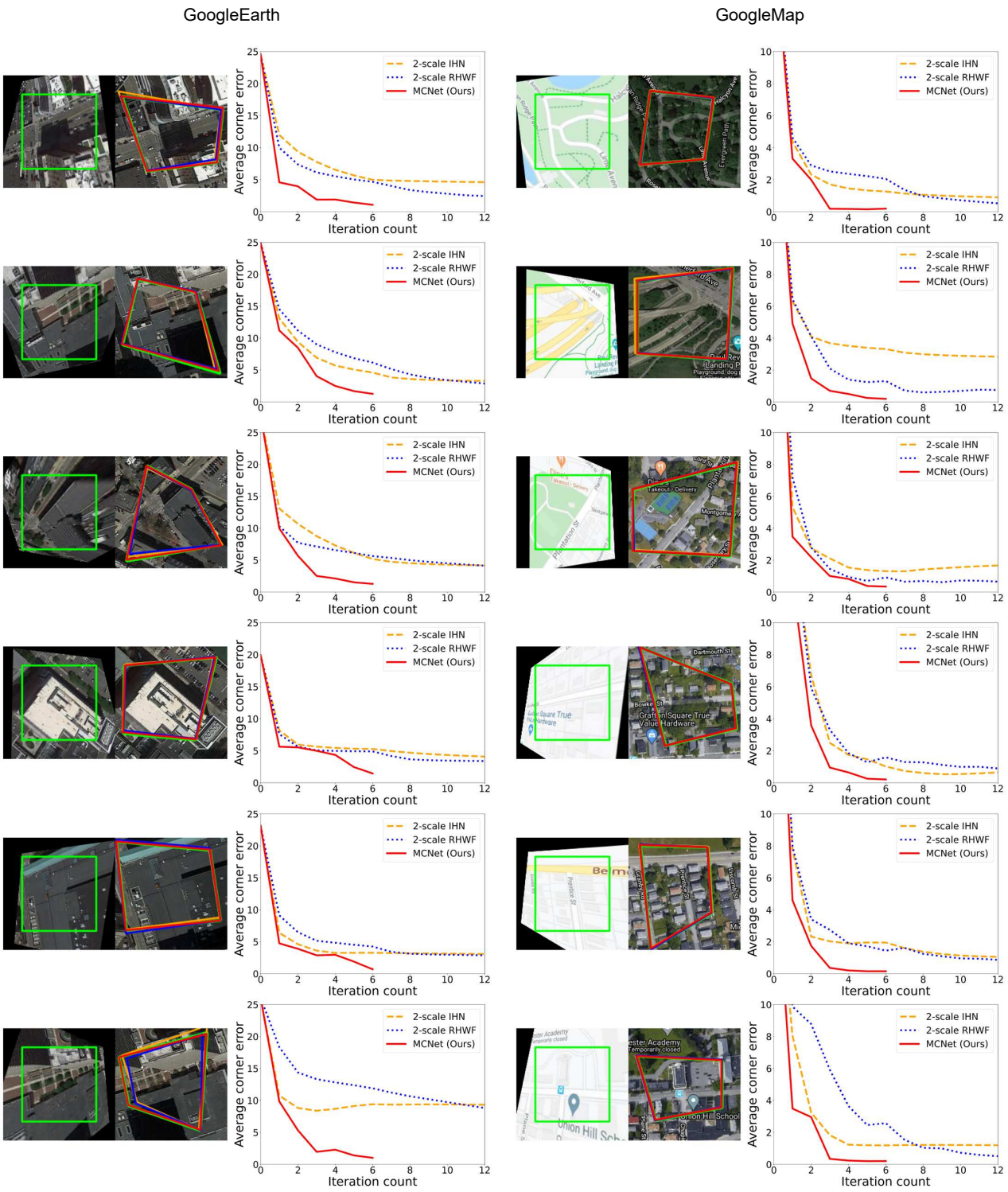

Figure 2. More experimental results of homography estimation with average corner error (ACE) at each iteration of our MCNet, IHN [2], and RHWF [3]. Green polygons denote the ground-truth position of $\mathbf{I}_{\mathrm{S}}$ on $\mathbf{I}_{\mathrm{T}}$. Red polygons denote the estimated position using our MCNet. Orange polygons denote the estimated position using 2-scale IHN. Blue polygons denote the estimated position using 2-scale RHWF. MCNet stops at iteration 6 while 2-scale IHN and 2-scale RHWF at 12.

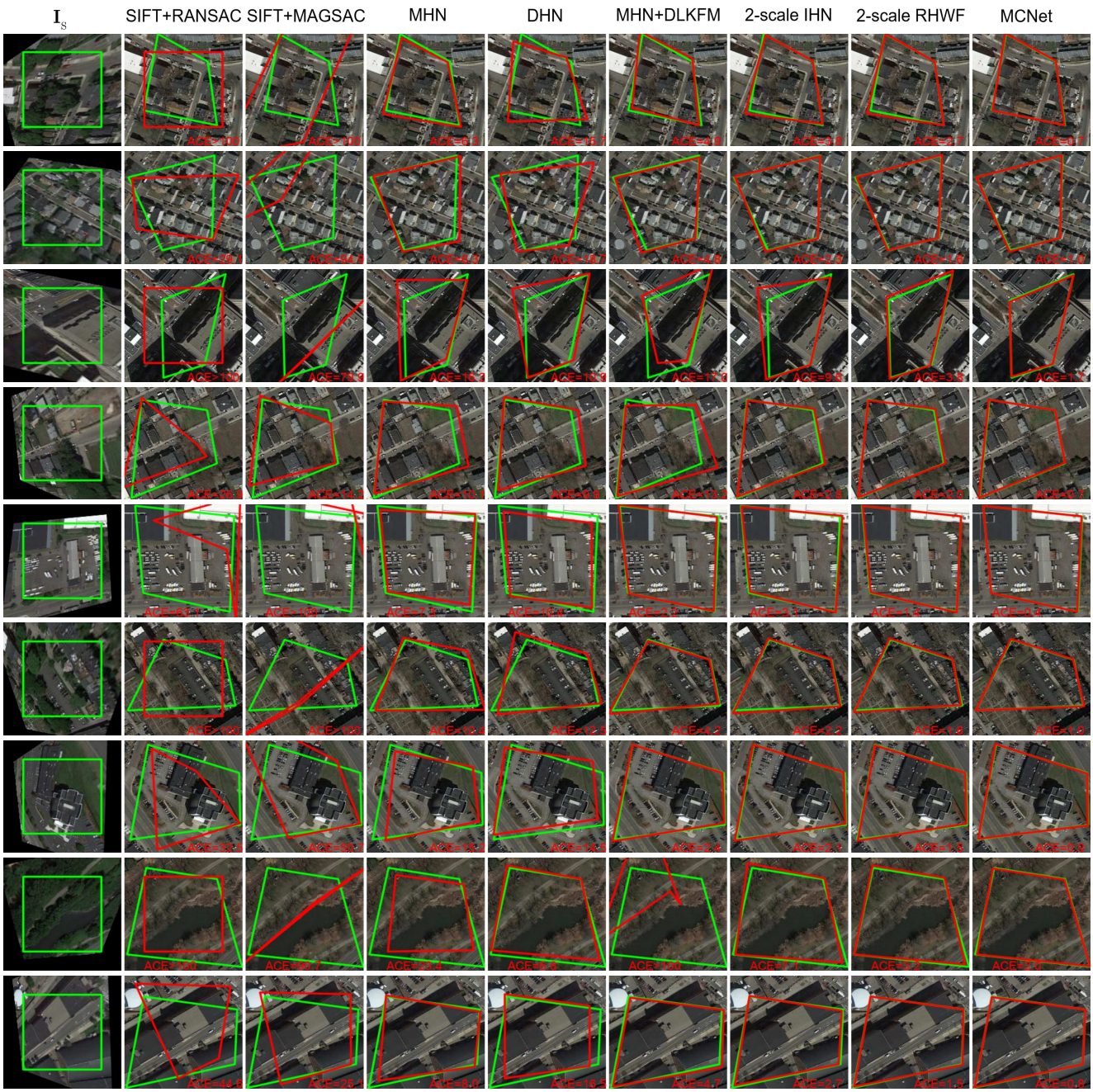

Figure 3. More experimental homography estimation results on the GoogleEarth dataset of various methods, including our MCNet, 2-scale RHWF [3], 2-scale IHN [2], MHN+DLKFM [11], DHN [5], MHN [6], SIFT+MAGSAC [1], and SIFT+RANSAC [8]. The green polygon represents the ground-truth location of the source image $\mathbf{I}_S$ on the target image. The red polygon represents the predicted location on the target image estimated by different algorithms. The smaller the relative distance of the polygons and the smaller the ACE the better estimation performance of the corresponding algorithm.

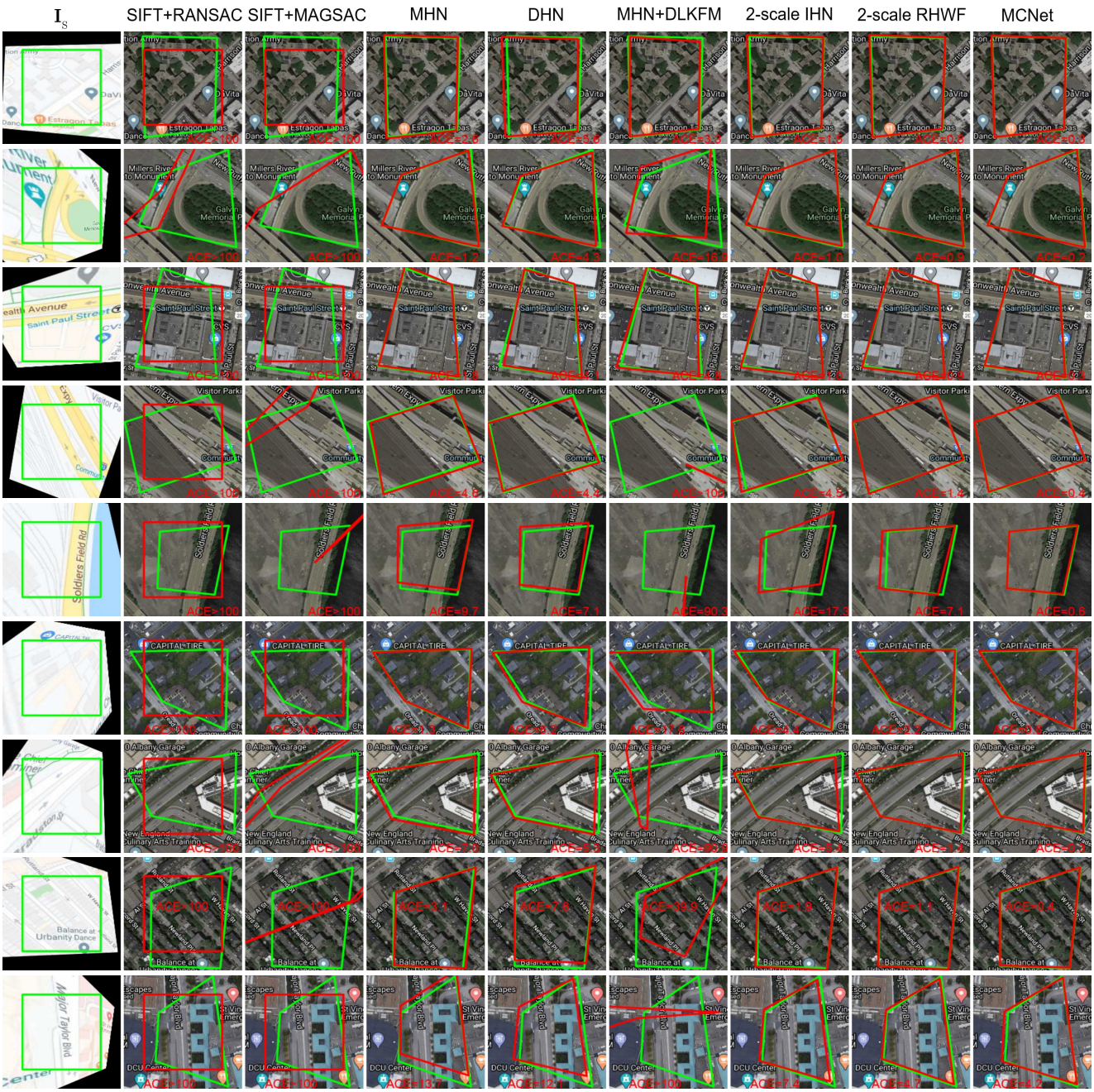

Figure 4. More experimental homography estimation results on the GoogleMap dataset of various methods, including our MCNet, 2-scale RHWF [3], 2-scale IHN [2], MHN+DLKFM [11], DHN [5], MHN [6], SIFT+MAGSAC [1], and SIFT+RANSAC [8]. The polygon settings are the same as in Fig. 3.

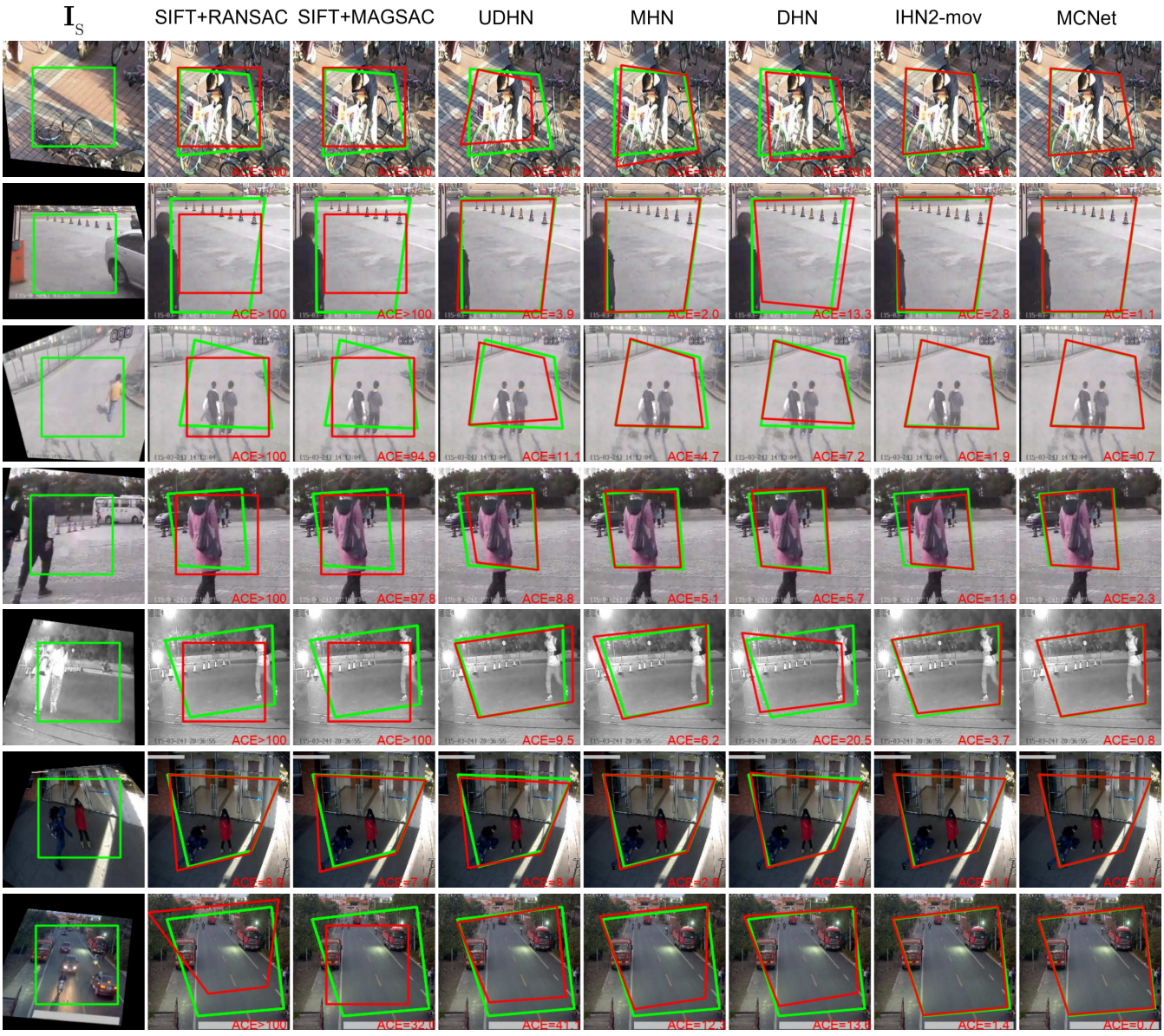

Figure 5. More experimental homography estimation results on the SPID dataset of various methods, including our MCNet, 2-scale IHN-mov [2], DHN [5], MHN [6], UDHN [10], SIFT+MAGSAC [1], and SIFT+RANSAC [8]. The polygon settings are the same as in Fig. 3.

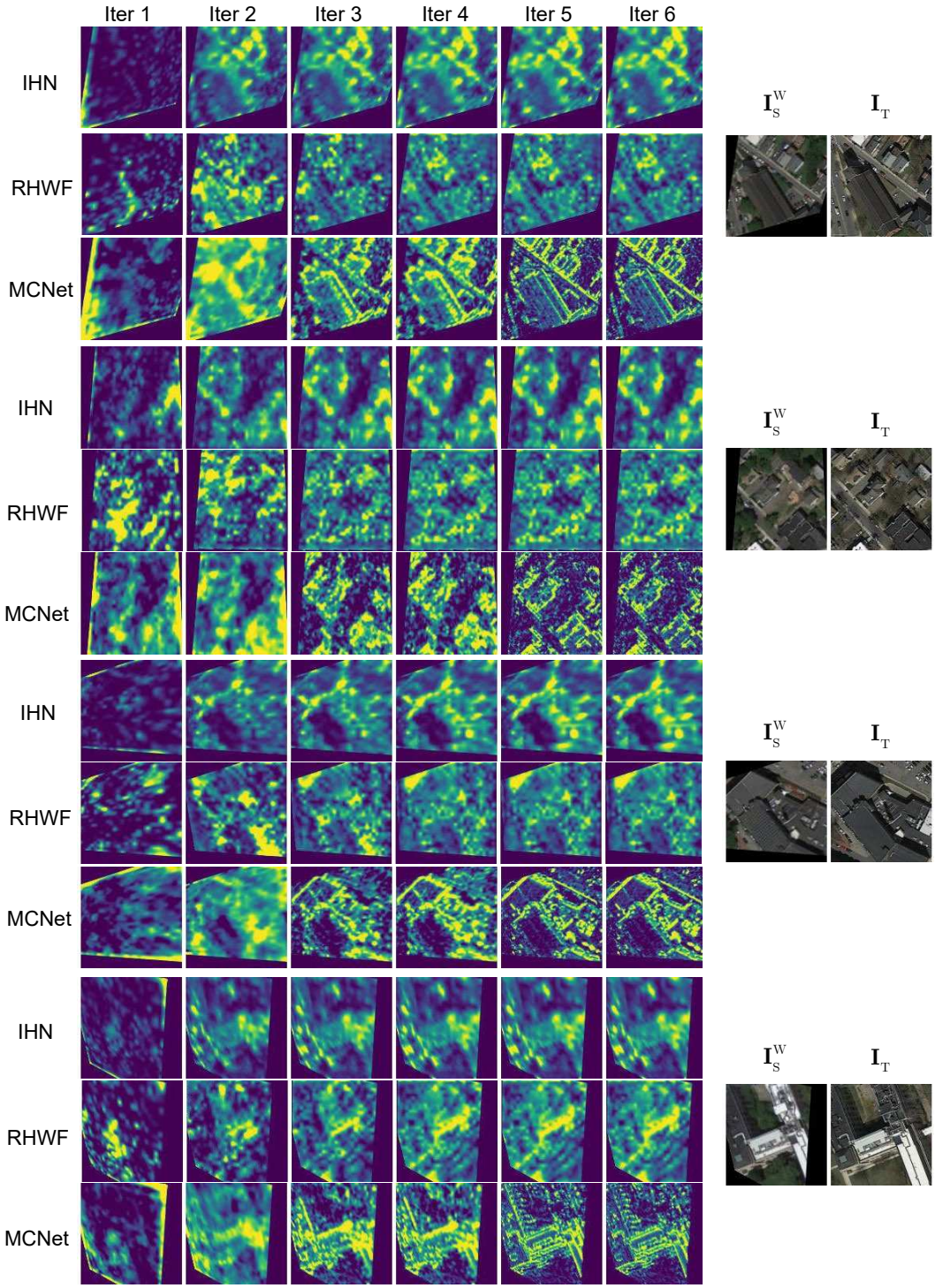

Figure 6. Comparison of correlation at each iteration for our MCNet, RHWF [3], and IHN [2] on the GoogleEarth dataset. The source image $\mathbf{I}_{\mathrm{S}}$ is warped to get $\mathbf{I}_{\mathrm{S}}^{\mathrm{W}}$, which is aligned with the target image $\mathbf{I}_{\mathrm{T}}$ to make a better illustration.

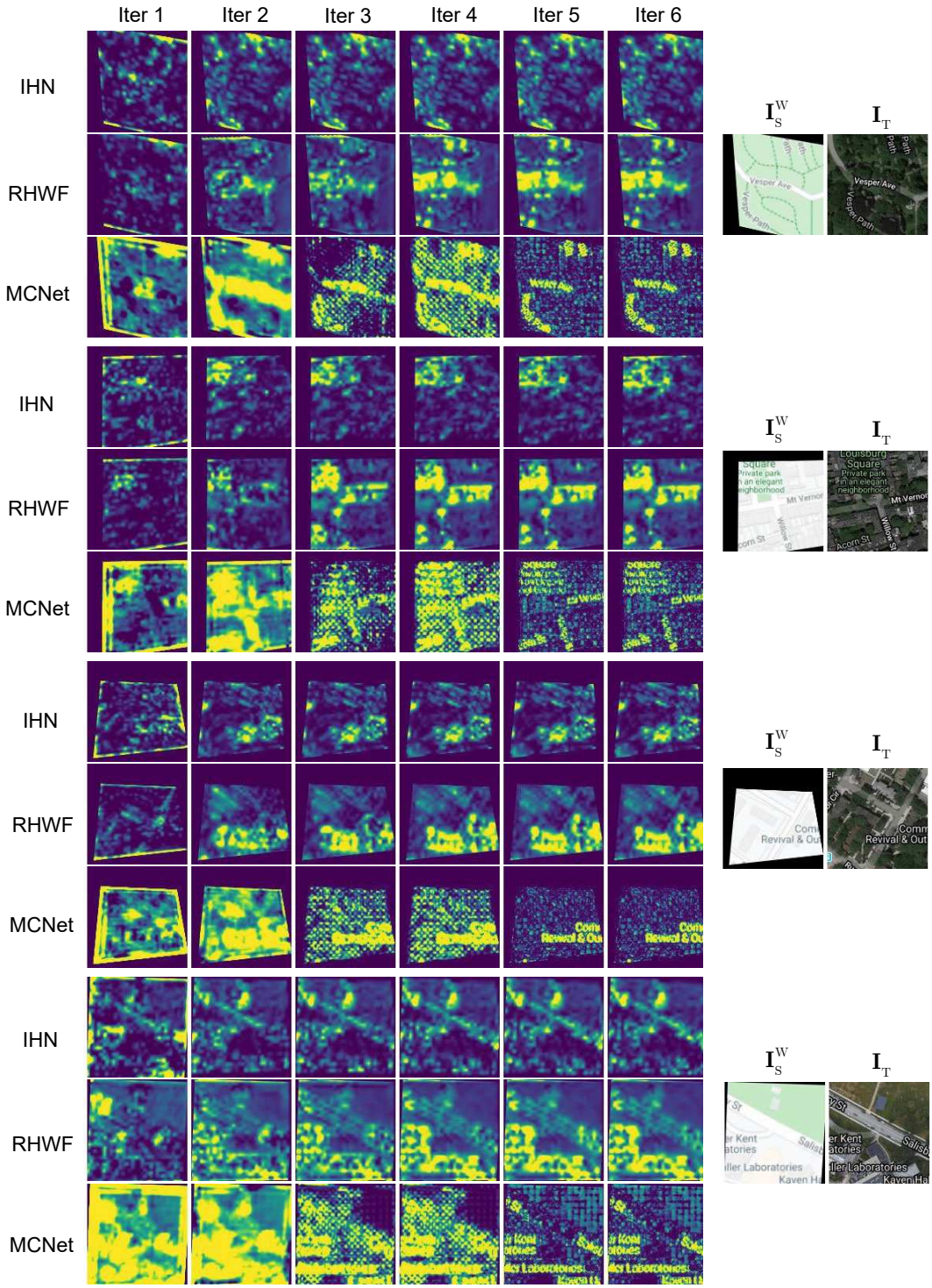

Figure 7. Comparison of correlation at each iteration for our MCNet, RHWF [3], and IHN [2] on the GoogleMap dataset. The image settings are the same as in Fig. 6.

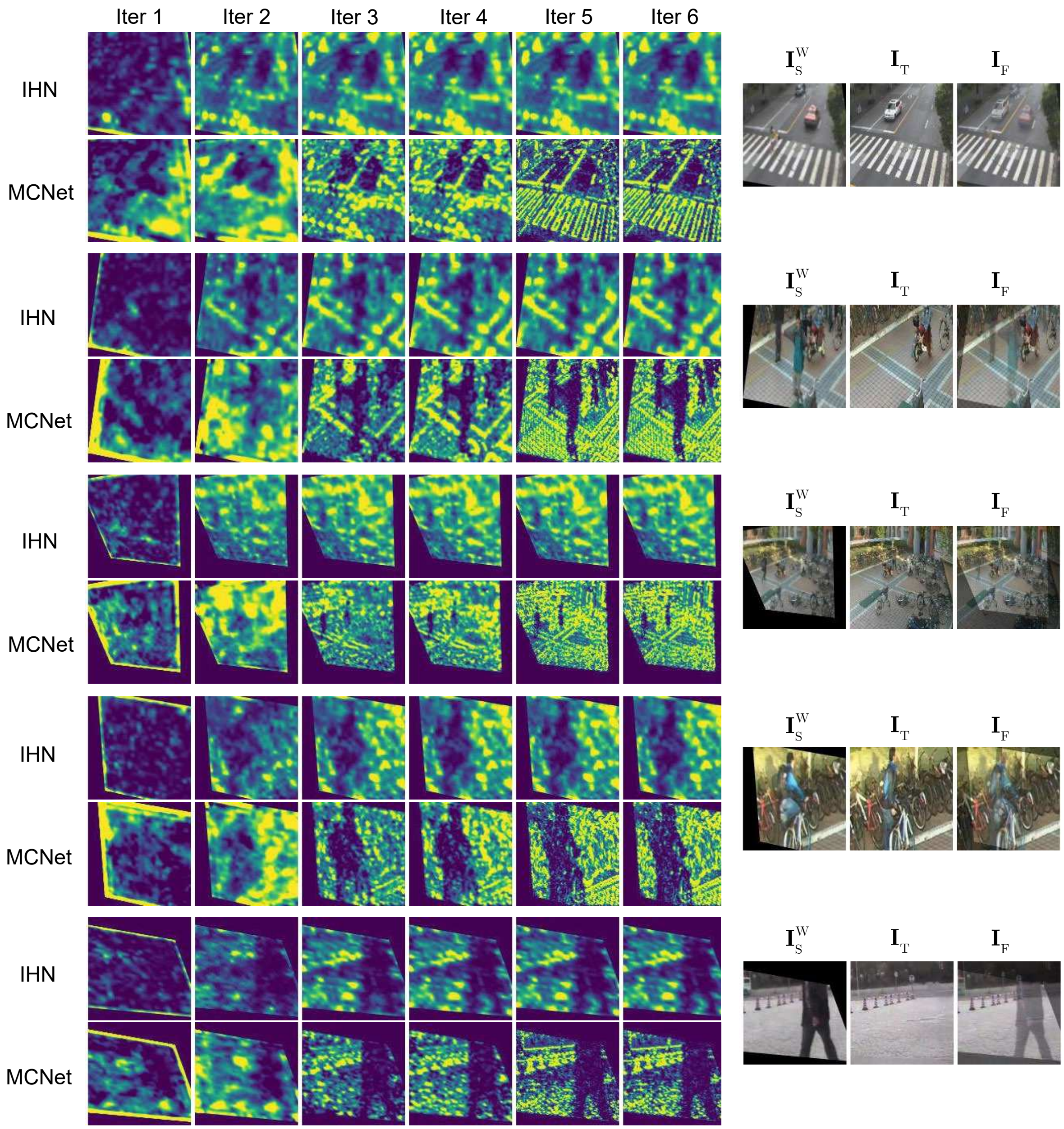

Figure 8. Comparison of correlation at each iteration for our MCNet and IHN [2] on the SPID dataset. The warped source image $\mathbf{I}_{\mathrm{S}}^{\mathrm{W}}$ and target image $\mathbf{I}_{\mathrm{T}}$ are fused to obtain $\mathbf{I}_{\mathrm{F}}$, in which the dynamic foreground objects are semitransparent.