Appendices

Due to space limitations, some experiments had to be omitted from the main paper. In this supplementary material, we show more comprehensive experimental results and evaluations.

Method	Run1	Run2	Run3	Run4
EdgeHomoNet	7.28	4.36	4.14	4.18
HomoNet	2.89	2.85	2.55	2.51

Table 1: Performance comparison (Mean Average Corner Error) of image-edge input network *vs*. image-image input network. The errors for identity matrix is 98.6.

A. Comparison with single-modality network

In this experiment, we train a baseline single-modality (image-to-image) homography network as [4]. Except for changing the network input from $[I_A, E_B]$ to $[I_A, I_B]$, all other factors are kept identical for the comparison. We refer to this baseline as HomoNet, and the network in the main paper as EdgeHomoNet. In both cases, the backbone network is ResNet-50. Note, to avoid the effect of image boundary during training, we apply a randomly generated mask on I_A before warping it to I_B based on perturbed 4 points.

Table 1 shows the performance comparison between EdgeHomoNet and HomoNet. As expected, the baseline HomoNet has better results given the texture and color information. This is consistent with results of [6]. However, our objective is not to replace the image to image (single-modality) homography estimation. Instead, we would like to address the problems of cross-modality with shape information, such as pixel-wise labeling for remote sensing images [10], multi-modal image alignment [12, 2], or sports field registration [3, 8, 9] *etc*. Interestingly, both HomoNet and EdgeHomoNet show more accurate results than the similar homography network experiments in [4]. This demonstrates the effectiveness of our training approaches.

B. Sports field registration experiments

B.1. Correlation layer

We also follow [5] in using correlation layer. Since we use two separate, non-identical processing streams (**A** and **B**) for input image I_A and edge image E_B, a *1x1* conv layer is added at the end of stream **A** to reduce the channels from 128 to 32. To limit the computation, a correlation layer with *11x11* window is used, as shown in Figure 1.

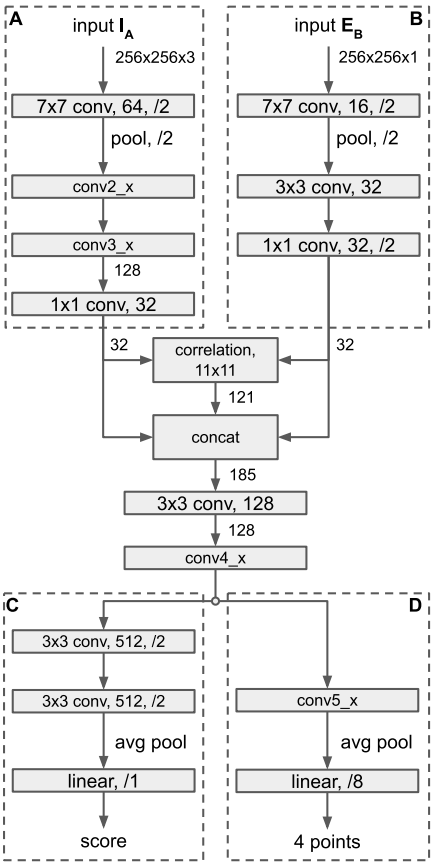


Figure 1: Example network architecture with *11x11* correlation layer. ResNet [7] is used as the backbone, and *conv2_x*, *conv3_x*, *conv4_x* and *conv5_x* are its four stacked building blocks.

B.2. Correlation layer evaluation on Hockey dataset

To evaluate the potential benefits of using a correlation layer, we propose to create more accurate hockey testing data for two reasons. Firstly, both WorldCup soccer dataset and Volleyball dataset have too few test samples. Secondly, we have observed large errors in WorldCup soccer ground truth. Thus, we create 10,000 ground truth images from 400 4K-resolution hockey images in the same way as the main paper. The 4K images are from two different arenas. In addition, because the evaluation with IoU_{part} and IoU_{whole} is highly sensitive to the field of view as observed in [11], we report mean average corner error defined in [4], which is actually average mean square errors over 4 points. We use the same training data and training process for both cases.

Table 2 shows that using a correlation layer leads to better performance on large testing data. Due to the computation cost related to the correlation layer, we only report the experiments without correlation layer in the main paper.

	Mean Average Corner Error
with correlation	**1.20**
w/o correlation	1.28

Table 2: Homography estimation comparison with and without correlation layer.

Method	Whole IoU		Part IoU	
	mean	median	mean	median
Original [9]	89.8	92.9	95.1	96.7
Corrected [9]	91.61	93.51	96.35	97.35
Original $Ours$	93.16	94.87	96.61	97.84
Corrected $Ours^1$	95.27	95.89	97.90	98.36
Corrected $Ours^2$	**96.06**	**96.98**	**98.28**	**98.78**

Table 3: Quantitative results for original ground truth and corrected ground truth on WorldCup soccer dataset. "Original" represents the ground truth from the original WorldCup dataset. "Corrected" is based on our manually corrected ground truth. For [9], we use its code [1] and network weights trained on original ground truth to perform evaluation. "Corrected $Ours^1$" represents the results trained with original ground truth and tested on corrected testing ground truth, while "Corrected $Ours^2$" trained with corrected ground truth.

B.3. WorldCup soccer dataset: visualization and ground truth errors

We have found that the original WorldCup dataset ground truth homographies show some errors. We illustrate all ground truths and test results with IoU_{part} less than 0.883 from our main paper test. As shown in Figure 2, all those cases with large errors are partially due to the ground truth errors.

To further demonstrate the effectiveness of our method, we manually correct the ground truth homographies (both training and testing data) and also report the results with the corrected ones. We will also include the corrected ground truth homographies in the supplementary materials, and will make them as well as the testing code publicly available in the future.

Table 3 shows the comparison for the results of original and corrected ground truth. For the state-of-the-art [9] method, we use its code and trained weights [1] to perform evaluation on both original and corrected testing data and observe improved performance on corrected ground truth. For our method, we also train a new model with corrected training data. Our results show a large improvement with almost perfect results.

References

[1] Vcg-uvic sports homography evaluation code. `https://github.com/vcg-uvic/sportsfield_release`. Accessed: 2020-10-11.

[2] Fereshteh S. Bashiri, Ahmadreza Baghaie, Reihaneh Rostami, Zeyun Yu, and Roshan D'Souza. Multi-modal medical image registration with full or partial data: A manifold learning approach. *Journal of Imaging*, 5:5, 12 2018.

[3] Jianhui Chen and James J. Little. Sports Camera Calibration via Synthetic Data. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

[4] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Deep image homography estimation. In *RSS Workshop on Limits and Potentials of Deep Learning in Robotics*, 2016.

[5] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. v. d. Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2758–2766, 2015.

[6] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *2019 International Conference on Learning Representations (ICLR)*, 2019.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *2016 European Conference on Computer Vision (ECCV)*, volume 9908, pages 630–645, 10 2016.

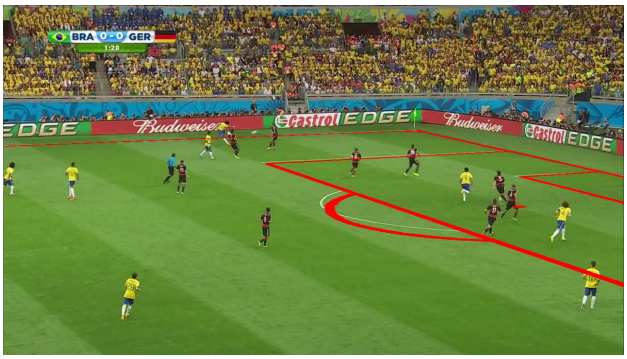
[8] Namdar Homayounfar, Sanja Fidler, and Raquel Urtasun. Sports field localization via deep structured models. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[9] Wei Jiang, Juan Camilo Gamboa Higuera, Baptiste Angles, Weiwei Sun, Mehrsan Javan, and Kwang Moo Yi. Optimizing through learned errors for accurate sports field registration. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.

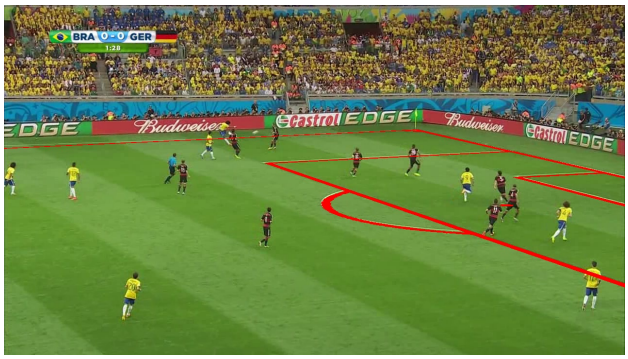
[10] E. Maggiori, G. Charpiat, Y. Tarabalka, and P. Alliez. Recurrent neural networks to correct satellite image classification maps. *IEEE Transactions on Geoscience and Remote Sensing*, 55(9):4962–4971, 9 2017.

[11] Long Sha, Jennifer Hobbs, Panna Felsen, Xinyu Wei, Patrick Lucey, and Sujoy Ganguly. End-to-end camera calibration for broadcast videos. In *2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13627–13636, 2020.

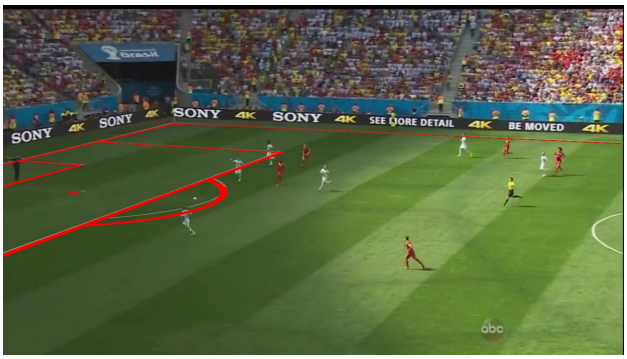
[12] Armand Zampieri, Guillaume Charpiat, Nicolas Girard, and Yuliya Tarabalka. Multimodal image alignment through a multiscale chain of neural networks with application to remote sensing. In *ECCV*, pages 679–696, 09 2018.



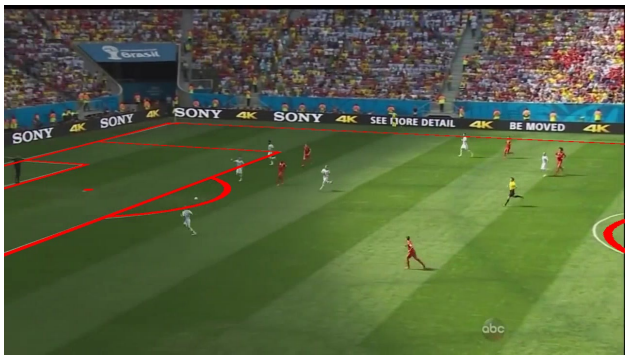
Ground truth



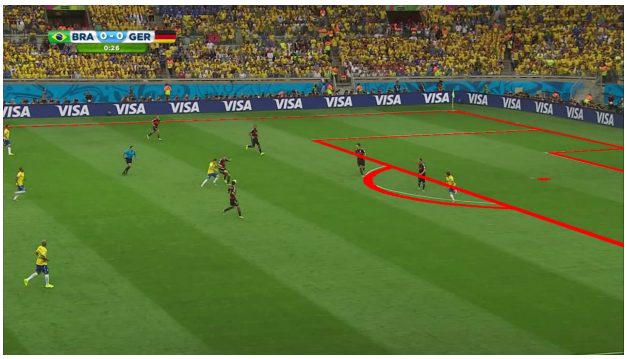
Results, IoU=0.819



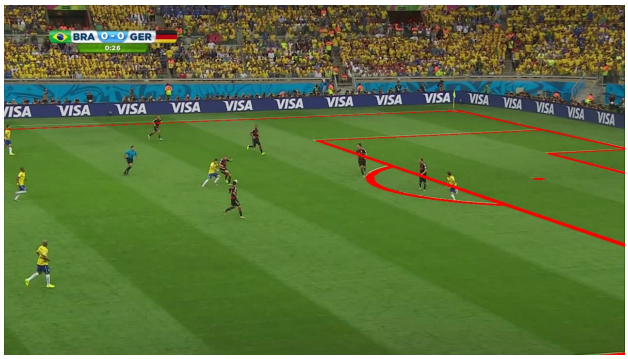
Ground truth



Results, IoU=0.84



Ground truth

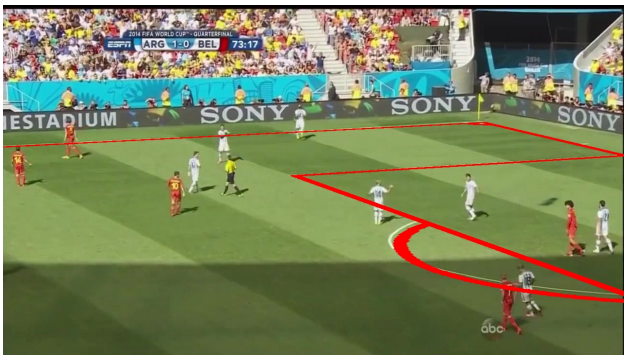


Results, IoU=0.859

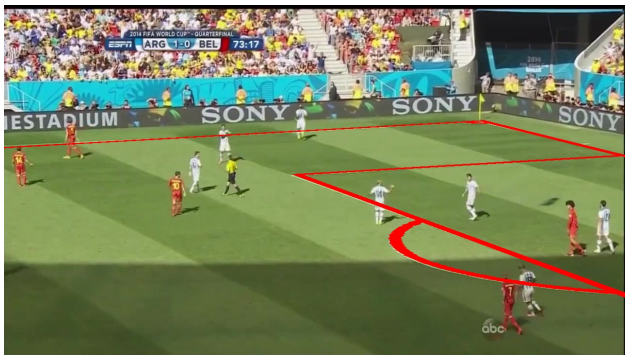
Ground truth

Results, IoU=0.864

Figure 2: Comparison of ground truths and our results with $IoU_{part} < 0.883$. Best viewed in color.



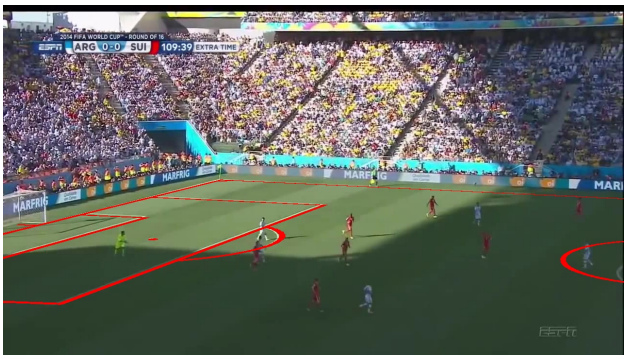
Ground truth



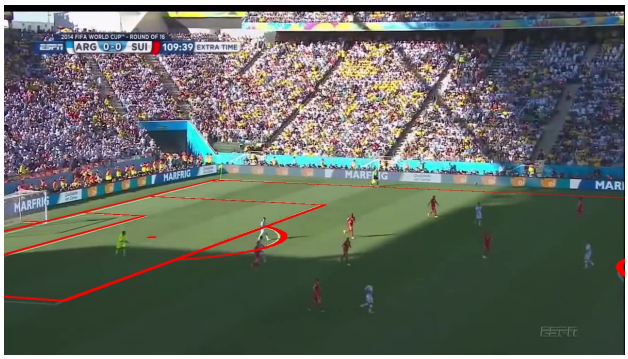
Results, IoU=0.861

Ground truth

Results, IoU=0.868



Ground truth



Results, IoU=0.874

Ground truth

Results, IoU=0.882

Figure 2: Comparison of ground truths and our results with $IoU_{part} < 0.883$. Best viewed in color.