Sports Field Registration via Keypoints-aware Label Condition

Yen-Jui Chu[1] Jheng-Wei Su[1] Kai-Wen Hsiao[1] Chi-Yu Lien[1] Shu-Ho Fan[1]
Min-Chun Hu[1] Ruen-Rone Lee[2] Chih-Yuan Yao[3] Hung-Kuo Chu[1]

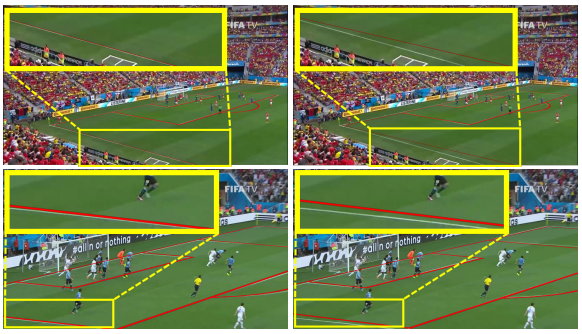
[1]National Tsing Hua University [2]Delta Research Center
[3]National Taiwan University of Science and Technology

Abstract

We propose a novel deep learning framework for sports field registration. The typical algorithmic flow for sports field registration involves extracting field-specific features (e.g., corners, lines, etc.) from field image and estimating the homography matrix between a 2D field template and the field image using the extracted features. Unlike previous methods that strive to extract sparse field features from field images with uniform appearance, we tackle the problem differently. First, we use a grid of uniformly distributed keypoints as our field-specific features to increase the likelihood of having sufficient field features under various camera poses. Then we formulate the keypoints detection problem as an instance segmentation with dynamic filter learning. In our model, the convolution filters are generated dynamically, conditioned on the field image and associated keypoint identity, thus improving the robustness of prediction results. To extensively evaluate our method, we introduce a new soccer dataset, called TS-WorldCup, with detailed field markings on 3812 time-sequence images from 43 videos of Soccer World Cup 2014 and 2018. The experimental results demonstrate that our method outperforms state-of-the-arts on the TS-WorldCup dataset in both quantitative and qualitative evaluations. Both the code and dataset are available online[1].

1. Introduction

Sports field registration refers to a process of estimating a homography transformation, which maps a 2D field template to a real-world field image captured from an arbitrary camera viewpoint. A robust sports field registration could benefit a wide variety of applications in sports broadcasting, including augmented sports tactics analysis, virtual advertisements insertion, true-view reply, etc. However, real-world field images usually present a uniform and textureless appearance, making the homography estimation a non-trivial and challenging task [16].



(a) Ours (b) Nie et al. [16]

Figure 1. **Visual comparisons of sports field registration.** We propose a novel keypoints-aware architecture (a) to obtain better field registration performance when comparing with the state-of-the-art method (b).

The typical algorithmic flow for sports field registration involves the following stages: (i) extracting field-specific features (e.g., corners, lines, circles, etc.) from field image; (ii) establishing the correspondence between the extracted features and the 2D field template; and (iii) estimating the homography matrix from the paired features using RANSAC [1]. The rapid development of deep learning has motivated several data-driven approaches for the prediction of field-specific features [3, 7, 12, 16] or regressing the homography matrix directly [14]. While recent works have shown impressive results, some still suffer from cases where the field images contain sparse features due to camera zoom-in or occlusions caused by the players. In light of this, Nie et al. [16] propose to detect a grid of uniformly distributed keypoints over the entire field. Such a dense sampling strategy increases the chances of predicting sufficient field features under various camera poses and thus leads to a more robust homography estimation.

Inspired by the work of Nie et al. [16], we use a grid

[1]https://ericsujw.github.io/KpSFR/

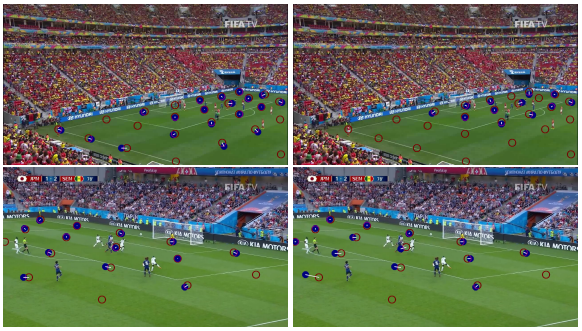
of uniformly distributed keypoints as our representation of field features. To detect keypoints, Nie et al. [16] associate each keypoint with a different class and train a semantic segmentation network. However, the uniform field appearance may easily confuse the network, causing the missing and misalignment problems during the keypoints prediction (see Figure 2(b)). In contrast, we propose to formulate the keypoints detection problem as an *instance* segmentation problem. We argue that detecting individual keypoints with similar visual content from a field image naturally aligns with the idea of instance segmentation. Specifically, we adopt the DoDNet [27] as the backbone model for the keypoints prediction. The dynamic filter learning mechanism of DoDNet enables the generation of adaptive kernels for each keypoint, which is effective for non-ROI-based instance segmentation [25]. Hence, our model can better retrieve keypoints with more accurate positions (see Figure 2(a)). As a result, our method obtains a better estimation of homography transformation as shown in Figure 1(a). To extensively evaluate our model and compare it with state-of-the-art methods, we introduce a new soccer dataset, called TS-WorldCup, by annotating time-sequence frames extracted from 43 videos of Soccer World Cup 2014 and 2018. The experiments demonstrate that our model achieves comparable results on the public WorldCup dataset and superior performance on the new TS-WorldCup dataset compared with several state-of-the-art methods.

Our contributions are as follows:

- We propose a novel deep neural network for sports field registration. Our model employs an instance segmentation architecture that leverages the dynamic filter learning to robustly predict a grid of uniformly distributed keypoints over the entire field image.
- We introduce a new soccer dataset, TS-WorldCup, with detailed field markings on 3812 field images, which is ten times larger than the public WorldCup dataset [12]. In addition, the TS-WorldCup dataset also contains time-sequence frames, which are beneficial for temporal evaluation.

2. Related Works

Sports field registration. Field registration is a critical component of most sports applications in computer vision [4, 19, 20, 22]. An intuitive approach is establishing the correspondences between images and the court field model and using the correspondences to estimate the homography. The correspondences generally are extracted with keypoint features [10, 15, 16] or higher-order features such as field lines [7, 8, 12]. For achieving more accurate registration, several works try to find a similar well-calibrated template as initial guessing for homography estimation. The well-calibrated templates are synthetic images with ground-truth homography transformation. The template images include only court field lines [3,24] or semantic of court area [6,23]. These features are not only for nearest-neighbor searching but for further homography refinement [14]. Different from existing approaches, our method builds the correspondence using densely sampled keypoints and adopts instance segmentation with *dynamic filter learning* for more robust and accurate registration.



(a) Ours (b) Nie et al. [16]

Figure 2. **Limitations of the state-of-the-art method.** The keypoints detection method by Nie et al. [16] may suffer the missing (top) and misalignment (bottom) problems due to the uniform field appearance. The groundtruth and predicted keypoints are denoted by hollow-red and solid-blue circles.

Dynamic filter learning. Dynamic filter learning is a recently popular mechanism to address the learned filters. In contrast, traditional convolutions are fixed for all samples. It has been introduced for achieving better network flexibility and representation capacity in lots of research field [13, 18, 25, 27]. Jia et al. [13] generate adaptive filters according to the input images to increase the flexibility of the network. They have also achieved impressive results in video and stereo prediction. On the other hand, Zhang et al. [27] train on partially labeled datasets with dynamic filter learning and have successfully and effectively segmented multiple organs and tumors. In this work, we demonstrate that *dynamic filter learning* can benefit field calibration by solving the ambiguity between distinct keypoints to improve the homography estimation accuracy.

3. Overview

Figure 3 illustrates the architecture overview, which consists of two main components: standard encoder-decoder and keypoints-aware label condition. The standard encoder-decoder architecture takes a field image I^{In} as input, encodes image features, and decodes the feature to a feature map (Section 4.1). In the keypoints-aware label condition module, we predict the final corresponding keypoints

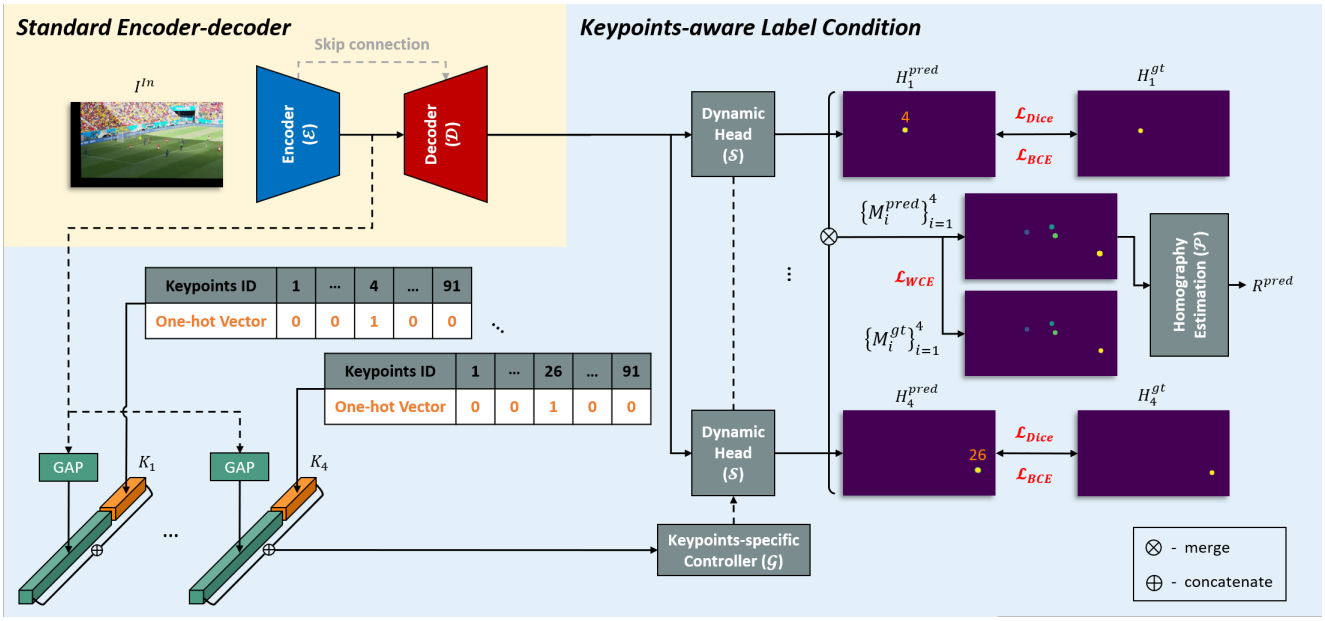


Figure 3. **Architecture overview.** Our proposed model consists of two parts, namely the standard encoder-decoder architecture and the keypoints-aware label condition module. Given the field image I^{In} as input, we perform symmetric encoder-decoder to extract the feature maps from encoder E and decoder D. We then generate the parameters of the dynamic head S using the keypoints-specific controller G fed by the extracted output feature of the encoder E (green vector) and the keypoints encoding vector K_i (orange vector). Then, the dynamic head S outputs the i-th heatmap H_i^{pred}. Finally, we employ soft aggregation used in [17] to merge all the predicted heatmaps $\{H_i^{pred}\}_{i=1}^{N}$ into the final output $\{M_i^{pred}\}_{i=1}^{N}$, and estimate the predicted homography R^{pred} using DLT [9] and RANSAC [1].

heatmap set $\{H_i^{pred}\}_{i=1}^{N}$ via dynamic head S using extracted features from the encoder E (green vector) and the assigned keypoints identity, which is encoded as a set of one-hot vectors $\{K_i\}_{i=1}^{N}$ (orange vector) where N is the number of keypoint class (Section 4.2). Then, we estimate the predicted homography R^{pred} from the merged heatmap $\{M_i^{pred}\}_{i=1}^{N}$, which is aggregated from the heatmap set $\{H_i^{pred}\}_{i=1}^{N}$. Finally, common loss functions in image segmentation, including the dice loss, the binary cross entropy loss, and the weighted cross entropy loss, are employed to train our model (Section 4.3).

4. Method

4.1. Standard Encoder-Decoder Architecture

We adopt a structure similar to U-Net [21], which is composed of one encoder E and one decoder D.

Encoder. The encoder E takes the field image I^{In} as the input and adopts ResNet-34 [11] as the backbone network, which incorporates dilated convolution [5] and non-local block layer [26].

Decoder. The decoder D uses four up-sampling blocks and skip connections to fuse encoder E features at different channels. The output feature map of the decoder D is bilinearly upsampled to the original resolution.

4.2. Keypoints-aware Label Condition

We pre-defined 91 uniformly distributed keypoints within the field template model and independently estimated each keypoint's heatmap during inference. For the traditional convolution networks, the learned filters are fixed during inference. However, we dynamically generate the convolution kernels conditioned on the current keypoints and the image feature extracted from the encoder E to predict heatmaps using keypoints-specific controller G and dynamic head S.

Dynamic filter generation. We encode the i-th keypoint as a 91-dimensional one-hot vector K_i. The image feature extracted by the encoder E is fed to a global average pooling (GAP) and then concatenated with keypoints encoding vector K_i as the input of the keypoints-specific controller G. Finally, the parameters of dynamic head S, i.e., containing weights and biases, are generated by keypoints-specific controller G and conditioned on both the field image I^{In} and the assigned keypoints.

Dynamic head. The dynamic head S takes the output of the keypoints-specific controller G and the output of the decoder D. It consists of three stacked convolution layers with 1×1 kernels. The first two layers have 16 channels, and the last layer has only one channel for final prediction. Therefore, a total of 561 parameters (weights: 528, biases: 33)

are generated by the keypoints-specific controller G. Finally, we employ soft aggregation used in Oh et al. [17] to merge all the predicted heatmaps $\{H_i^{pred}\}_{i=1}^{N}$ into the final output $\{M_i^{pred}\}_{i=1}^{N}$ and estimate the predicted homography R^{pred} using DLT [9] and RANSAC [1].

4.3. Loss Functions

Binary dice loss. Dice loss is proven helpful for addressing the data imbalance problem between foreground and background. Its formulation is as follows:

$$\begin{aligned} L_{Dice} &= 1 - \text{Dice Coefficient} \\ \text{Dice Coefficient} &= \frac{2 \mid H_i^{gt} \bigcap H_i^{pred} \mid}{\mid H_i^{gt} \mid + \mid H_i^{pred} \mid} \end{aligned} \tag{1}$$

where H_i^{gt} is ith channel of ground-truth keypoints heatmap, and H_i^{pred} is the ith channel of predicted keypoints heatmap of the field image I^{In}. $\mid H_i^{gt} \bigcap H_i^{pred} \mid$ represents the intersection of the H_i^{gt} set, the H_i^{pred} set, and the denominator represents the number of corresponding elements.

Binary cross entropy (BCE) loss. BCE loss is commonly used in the binary classification problems. It compares each of the predicted probabilities to actual class output. We define the BCE loss as follows:

$$\begin{aligned} L_{BCE} = &- \sum_{i}^{Q} H_i^{gt} \cdot \log\left(\sigma\left(H_i^{pred}\right)\right) \\ &+ (1 - H_i^{gt}) \cdot \log\left(\sigma\left(1 - H_i^{pred}\right)\right) \end{aligned} \tag{2}$$

where σ is the sigmoid function and Q is the number of keypoints we randomly selected.

Weighted cross entropy (WCE) loss. WCE loss tackles the data imbalance problem by assigning weight to each class. We define the BCE loss as follows:

$$L_{WCE} = -\sum_{i}^{Q} w_i \cdot M_i^{gt} \cdot \log\left(M_i^{pred}\right) \tag{3}$$

where M_i^{gt} is the ground-truth keypoints heatmap, M_i^{pred} is the corresponding predicted keypoints heatmap merged by soft aggregation [17]. The first channel indicates background, and w_i is the corresponding weight. Q is the number of keypoints we randomly selected.

Total loss. The overall loss function is defined as follows:

$$L_{total} = \lambda_{Dice} L_{Dice} + \lambda_{BCE} L_{BCE} + \lambda_{WCE} L_{WCE} \tag{4}$$

where λ_{Dice}, λ_{BCE}, and λ_{WCE} are the hyperparameters for weighting the loss functions.

5. TS-WorldCup Dataset

Considering the insufficient among of field images in the original WorldCup dataset [12], we create a new soccer dataset with detailed field markings on 3812 field images from 43 videos of Soccer World Cup 2014 and 2018, which is ten times larger than the WorldCup dataset. Since the dataset contains time-sequence information, we call this dataset the TS-WorldCup dataset. In the next section, we evaluate the performance of our method on TS-WorldCup with the data splits of 2925 images for training and 887 images for testing.

6. Experiments

6.1. Experimental Settings

Baselines. We compare our model with the following state-of-the-art sports field registration methods:

- Homayounfar et al. [12] is a learning-based method, which leverages field features such as circles and lines and formulates the field registration problem using a Markov Random Field during inference.
- Sharma et al. [24] is a fully automatic and template matching-based method, which leverages edge information instead of point correspondence and synthetic data to enhance camera registration accuracy.
- Sha et al. [23] is a template matching-based method using an end-to-end architecture to estimate camera calibration.
- Jiang et al. [14] is an optimization-based method leveraging the model to predict the registration error and optimize the predicted homography.
- Citraro et al. [7] is a learning-based method training on the WorldCup dataset with additional players' location annotation. On the other hand, we follow the setting from Nie et al. [16] and report the results without using extra annotation for a fair comparison.
- Cioppa et al. [6] is a learning-based method training on the SoccerNet dataset [6] with the pseudo-ground-truth calibrations generated by Xeebra [2].
- Xeebra [2] is a commercial multi-camera review system using a machine learning method to predict the camera registration on the soccer field without fine-tuning on the WorldCup dataset.
- Chen et al. [3] is a template matching-based method. We use the data provided by the author (obtained from the WorldCup dataset) to train the two-GAN model, and the provided 90K feature-pose database is used to train the siamese network.

Table 1. **Quantitative comparisons on the WorldCup dataset.** The symbols, † and ⋆ denote respectively not fine-tunning on WorldCup dataset and our reimplementation of Nie et al. [16].

Method	$IOU_{entire}(\%)\uparrow$		$IOU_{part}(\%)\uparrow$		Proj.(meter) ↓		Re-Proj. ↓	
	mean	medium	mean	medium	mean	medium	mean	medium
Homayounfar et al. [12]	83	-	-	-	-	-	-	-
Sharma et al. [24]	-	-	91.4	92.7	-	-	-	-
Sha et al. [23]	88.3	92.1	93.2	96.1	-	-	-	-
Jiang et al. [14]	89.8	92.9	95.1	96.7	1.21	0.74	**0.017**	**0.012**
Citraro et al. [7]	90.5	91.8	-	-	-	-	0.018	**0.012**
Cioppa et al. [6]†	79.8	81.7	88.5	92.3	-	-	-	-
Xeebra [2]†	91.7	93	**96.7**	**98.7**	-	-	-	-
Chen et al. [3]	89.4	**93.8**	94.5	96.1	-	-	-	-
Nie et al. [16] - keypoints Only	91.5	93.3	95.8	97.2	0.82	0.61	0.019	0.015
Nie et al. [16] - alignment	**91.6**	93.4	95.9	97.1	0.84	0.65	0.019	0.014
Nie et al. [16]⋆	90.3	92.2	95.9	97.0	0.82	**0.60**	0.018	0.015
Ours	91.2	93.1	96.0	97.0	**0.81**	0.63	0.019	0.014

- Nie et al. [16] is an optimization-based method. Since there are no publicly available codes from the authors, we re-implement the network prediction module (keypoints only) and post-processing steps.

Evaluation metrics. We take three commonly used metrics in quantitative evaluation, including intersection over union (IoU), projection error, and re-projection error. For each metric, we report both mean and median:

- **IoU.** We compute the IOU_{part} and IOU_{entire} by comparing the two binary masks projected from predicted homography R^{pred} and ground-truth homography R^{gt}.
- **Projection error.** The projection error is computed by the average distance in actual scale (meters) between projected points using predicted and ground-truth homography. To calculate the distance, uniformly sample 2500 pixels from the visible field area of the camera image and project them to the field. The actual field dimension for soccer is 100×60.
- **Re-projection error.** The re-projection error is computed using predicted and ground-truth homography by the average distance between points re-projected in the video frame .

Implementation details. We conduct the experiments on a single NVIDIA V100 with 32G VRAM and implement our model in PyTorch. We use Adam as the optimizer with the learning rate of 1e-4 and batch size of 4. We use cross entropy loss with the weight of each keypoint set to 100 while background pixels are set to 1 and binary cross entropy loss with $Q = 4$. We empirically set $\lambda_{Dice} = 1$, $\lambda_{BCE} = 1$, and $\lambda_{WCE} = 1$ in the total loss function (Equation 4).

During the training process of the WorldCup dataset, we train our model for 1500 epochs, and the scheduler decays the learning rate by 0.1 every 300 epochs. In the fine-tuning process of our TS-WorldCup, we train our model for 100 epochs, and the scheduler decays the learning rate by 0.1 every 30 epochs. As for the inference process, we decode predicted heatmaps $\{M_i^{pred}\}_{i=1}^{N}$ into keypoint sets using Non-Maximum Suppression (NMS) to get the keypoint positions and then find correspondences from field model, followed by estimating homography transformation using RANSAC [1] and DLT [9]. The re-projection threshold for RANSAC [1] is 10.

6.2. Evaluation on WorldCup Dataset

In this experiment, we evaluate the performance of our model quantitatively by comparing with baselines on the WorldCup dataset. Note that since there are no publicly available inference codes for some baselines [2, 6, 7, 12, 23, 24], we directly report the statistics from their papers. As shown in Table 1, our method is comparable with baselines on several metrics, especially on IOU_{entire} and projection error comparing with Jiang et al. [14], Chen et al. [3] and Nie et al. [16]. However, the WorldCup dataset contains only 186 field images in its testing set, which is small and may cause bias. In the following section, we evaluate our method on our TS-WorldCup dataset.

6.3. Evaluation on TS-WorldCup Dataset

This experiment compares our method with baselines quantitatively and qualitatively on the TS-WorldCup

Table 2. **Quantitative comparisons on the TS-WorldCup dataset.** The methods in the first block are trained using the WorldCup dataset. The symbol $*$ denotes the methods that are finetuned on the TS-WorldCup training set.

Method	$IOU_{entire}(\%)\uparrow$		$IOU_{part}(\%)\uparrow$		Proj.(meter) $\downarrow$		Re-Proj. $\downarrow$	
	mean	medium	mean	medium	mean	medium	mean	medium
Jiang et al. [14]	88.1	90.7	94.8	95.0	1.07	0.91	0.040	0.040
Chen et al. [3]	89.0	92.2	96.8	97.6	0.65	0.47	0.020	0.017
Nie et al. [16]	90.1	92.8	96.6	97.4	0.57	0.51	0.015	0.012
Ours	**93.2**	**94.3**	**97.6**	**97.7**	**0.45**	**0.41**	**0.012**	**0.011**
Chen et al. [3]*	90.7	94.1	96.8	97.4	0.54	0.38	0.016	0.013
Nie et al. [16]*	92.5	94.2	97.4	97.8	0.43	0.38	0.011	0.010
Ours*	**94.8**	**95.4**	**98.1**	**98.2**	**0.36**	**0.33**	**0.009**	**0.008**

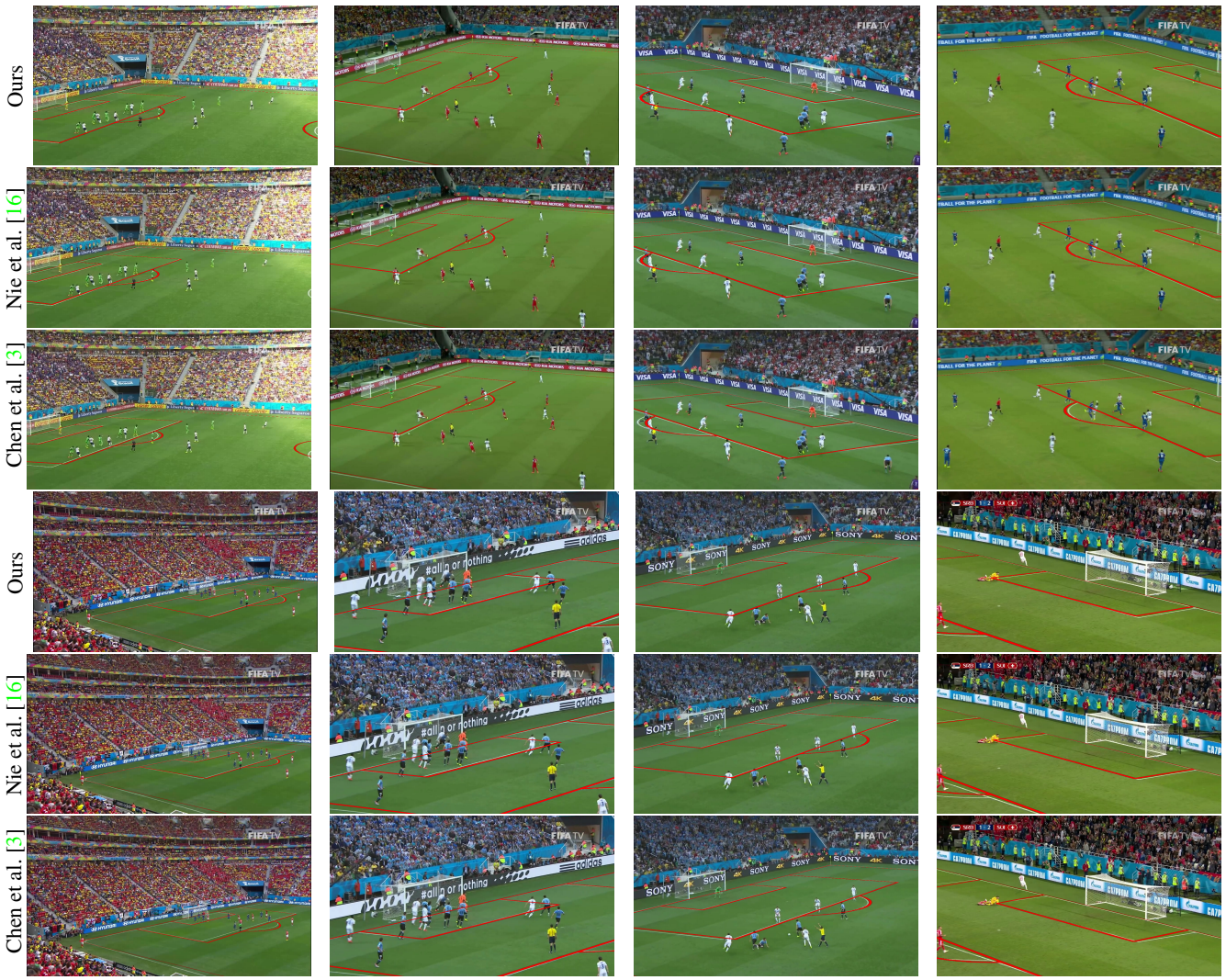


Figure 4. **Qualitative comparisons on the TS-WorldCup dataset.** Our method estimates better registration results compared with other competing methods.

Table 3. **Quantitative results of the ablation study.** We evaluate the effectiveness of the different loss functions in our model on TS-WorldCup dataset.

Dice	BCE	WCE	$IOU_{entire}(\%)\uparrow$		$IOU_{part}(\%)\uparrow$		Proj.(meter) $\downarrow$		Re-Proj. $\downarrow$	
			mean	medium	mean	medium	mean	medium	mean	medium
✓	✓	×	71.7	81.5	87.2	92.7	2.22	1.28	0.044	0.026
✓	×	✓	91.3	94.1	96.1	97.5	0.54	**0.41**	0.019	0.012
✓	✓	✓	**93.2**	**94.3**	**97.6**	**97.7**	**0.45**	**0.41**	**0.012**	**0.011**

Table 4. **Quantitative results on keypoints detection.** The symbol $*$ denotes the model finetuned on our TS-WorldCup dataset.

Method	MSE $\downarrow$		Recall $\uparrow$
	common	complete	
Nie et al. [16]	2.41	2.43	0.77
Ours	**2.02**	**2.3**	**0.82**
Nie et al. [16]*	1.86	1.92	0.83
Ours*	**1.54**	**1.7**	**0.87**

dataset. The experiment contains two parts. First, for all the methods, we use the pre-trained weights on the WorldCup dataset and test on the TS-WorldCup testing set (see the upper block in Table 2). As for the evaluation on Jiang et al. [14], we report projection error and re-projection error in this part, since we evaluate all the methods here by our own implementation. Second, we fine-tune all the methods except Jiang et al. [14] (without official training code) on the TS-WorldCup training set and test on the TS-WorldCup testing set (see the lower block in Table 2). In general, our method achieves the best performance against all the baselines across all evaluation metrics with or without fine-tuning. As shown in Figure 4, our method clearly outperforms baselines in estimating the soccer field registration that aligns the field line well across many different circumstances. Please refer to our online webpage for more qualitative comparisons[2].

6.4. Comparisons on Keypoints Detection

Since we argue that our instance-based keypoints detection architecture can better retrieve individual keypoints with more accurate positions, we further conducted a quantitative evaluation to prove this point. In Table 4, we report the mean square error (MSE) on i) the intersection of keypoints predicted from our method and baseline (common); and ii) all the predicted keypoints from each method (complete). We also calculate the average recall rate of keypoints for both methods. We can tell that our instance-based keypoints detection achieves better accuracy and recall rate against the baseline.

6.5. Ablation Study

Here, we conduct ablation studies to validate the effectiveness of the different loss functions in our model. We combine binary dice (Dice) loss with binary cross-entropy (BCE) loss or weighted cross-entropy (WCE) loss, respectively. As shown in Table 3, without using WCE loss, our model suffers a significant performance drop in all metrics. This indicates that the large background area biases the keypoints prediction toward the background class. The influence of removing BCE loss is mild, and we obtain the best performance in the model trained with all loss functions. Please refer to our online webpage for more qualitative comparisons[2].

7. Conclusions

In this paper, we propose a novel framework for sports field registration. We incorporate dynamic filter learning to generate kernels dynamically for each keypoint. We have created a large new soccer dataset with time sequence, termed TS-WorldCup, from Soccer World Cup 2014 and 2018, almost ten times the public WorldCup dataset [12]. Our proposed method achieves promising results and further proves that sports field registration tasks can benefit from dynamic filter learning. In the future, we aim to extend our approach to other sports such as basketball and baseball.

Acknowledgements. The project was funded in part by the Ministry of Science and Technology of Taiwan (108-2221-E-011-097-MY3, 108-2221-E-007-106-MY3, 109-2221-E-007-095-MY3, 110-2221-E-007-061-MY3, and 110-2221-E-007-060-MY3).

References

[1] Opencv: Camera calibration and 3d reconstruction. https://docs.opencv.org/3.4.11/d9/d0c/group__calib3d.html#

[2] https://ericsujw.github.io/KpSFR/

ga4abc2ece9fab9398f2e560d53c8c9780, 2021. Online; accessed: 2021-12-22. 1, 3, 4, 5

[2] Multi-camera review system - xeebra | evs. https://evs.com/products/video-assistance/xeebra, 2022. Online; accessed: 2022-03-14. 4, 5

[3] Jianhui Chen and James J Little. Sports camera calibration via synthetic data. In *CVPRW*, 2019. 1, 2, 4, 5, 6

[4] Jun Chen, Ryosuke Watanabe, Keisuke Nonaka, Tomoaki Konno, Hiroshi Sankoh, and Sei Naito. Fast free-viewpoint video synthesis algorithm for sports scenes. In *IROS*, 2019. 2

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 3

[6] Anthony Cioppa, Adrien Deliege, Floriane Magera, Silvio Giancola, Olivier Barnich, Bernard Ghanem, and Marc Van Droogenbroeck. Camera calibration and player localization in soccernet-v2 and investigation of their representations for action spotting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021. 2, 4, 5

[7] Leonardo Citraro, Pablo Márquez-Neila, Stefano Savare, Vivek Jayaram, Charles Dubout, Félix Renaut, Andrés Hasfura, Horesh Ben Shitrit, and Pascal Fua. Real-time camera pose estimation for sports fields. *Machine Vision and Applications*, 2020. 1, 2, 4, 5

[8] Carlos Cuevas, Daniel Quilon, and Narciso García. Automatic soccer field of play registration. *Pattern Recognition*, 2020. 2

[9] RI Hartleyand and A Zisserrnan. Multiple view geometry, 2004. 3, 4, 5

[10] Jean-Bernard Hayet, Justus Piater, and Jacques Verly. Robust incremental rectification of sports video sequences. In *BMVC*, 2004. 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[12] Namdar Homayounfar, Sanja Fidler, and Raquel Urtasun. Sports field localization via deep structured models. In *CVPR*, 2017. 1, 2, 4, 5, 7

[13] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. *NIPS*, 2016. 2

[14] 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 *WACV*, 2020. 1, 2, 4, 5, 6, 7

[15] Jikai Lu, Jianhui Chen, and James J Little. Pan-tilt-zoom slam for sports videos. In *BMVC*, 2019. 2

[16] Xiaohan Nie, Shixing Chen, and Raffay Hamid. A robust and efficient framework for sports-field registration. In *WACV*, 2021. 1, 2, 4, 5, 6, 7

[17] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, 2019. 3, 4

[18] Youwei Pang, Lihe Zhang, Xiaoqi Zhao, and Huchuan Lu. Hierarchical dynamic filtering network for rgb-d salient object detection. In *ECCV*, 2020. 2

[19] Konstantinos Rematas. Watching sports in augmented reality. *IEEE Potentials*, 2019. 2

[20] Konstantinos Rematas, Ira Kemelmacher-Shlizerman, Brian Curless, and Steve Seitz. Soccer on your tabletop. In *CVPR*, 2018. 2

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 3

[22] Hiroshi Sankoh, Sei Naito, Keisuke Nonaka, Houari Sabirin, and Jun Chen. Robust billboard-based, free-viewpoint video synthesis algorithm to overcome occlusions under challenging outdoor sport scenes. In *ACM MM*, 2018. 2

[23] Long Sha, Jennifer Hobbs, Panna Felsen, Xinyu Wei, Patrick Lucey, and Sujoy Ganguly. End-to-end camera calibration for broadcast videos. In *CVPR*, 2020. 2, 4, 5

[24] Rahul Anand Sharma, Bharath Bhat, Vineet Gandhi, and CV Jawahar. Automated top view registration of broadcast football videos. In *WACV*, 2018. 2, 4, 5

[25] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *ECCV*, 2020. 2

[26] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 3

[27] Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen. Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In *CVPR*, 2021. 2