A Robust and Efficient Framework for Sports-Field Registration

Xiaohan Nie, Shixing Chen, and Raffay Hamid
Amazon Prime Video
{nxiaohan, shixic, raffay}@amazon.com

Abstract

We propose a novel framework to register sports-fields as they appear in broadcast sports videos. Unlike previous approaches, we particularly address the challenge of field-registration when: (a) there are not enough distinguishable features on the field, and (b) no prior knowledge is available about the camera. To this end, we detect a grid of keypoints distributed uniformly on the entire field instead of using only sparse local corners and line intersections, thereby extending the keypoint coverage to the texture-less parts of the field as well. To further improve keypoint based homography estimate, we differentialbly warp and align it with a set of dense field-features defined as normalized distance-map of pixels to their nearest lines and key-regions. We predict the keypoints and dense field-features simultaneously using a multi-task deep network to achieve computational efficiency. To have a comprehensive evaluation, we have compiled a new dataset called SportsFields which is collected from 192 video-clips from 5 different sports covering large environmental and camera variations. We empirically demonstrate that our algorithm not only achieves state of the art field-registration accuracy but also runs in real-time for HD resolution videos using commodity hardware.

1. Introduction

Sports-field registration requires mapping image-pixels of a sports-field to their corresponding real-world locations. It is an important step towards building perceptual understanding of sports and enables multiple sports-related applications including novel view-point synthesis, virtual ads-placement in and around the field, and generation of multiple sports-visualizations informative to sports audience.

The standard approach for field registration consists of: (i) extracting salient keypoints on the field-image [32, 14, 25], (ii) finding the pair-wise correspondences between field-image and a field-template by matching the local descriptors [27, 2] of the extracted keypoints, and (iii) estimating the image-to-template transformation which is usually parameterized as a homography matrix [12, 23, 15]. Recent approaches have often used deep learning to estimate this homography matrix either by applying a deep network to predict the field markings which are then used to estimate the homography [17, 5, 9, 34], or regressing the homography parameters directly [19, 28, 10, 22]. While these methods offer promising results, having an approach that accurately generalizes to a large set of sports remains unsolved.

Figure 1. Example frames illustrating three main factors that result in lack of sufficient distinct field-features: (a – row 1) uniform and texture-less field appearance, (b – row 2) narrow field of view, and (c – row 3) field-occlusions due to presence of multiple players.

Table 1 summarizes the main factors that result in making field-registration a challenging problem for a set of five different sports. The interplay among these factors poses various technical challenges for sports-field registration, the most important of which is the lack of sufficient distinct field-features (see Figure 1 for some illustrative examples). In particular, this challenge emerges from:

a. Uniform field-appearance: in American football, the field looks like a chessboard with lots of identical line intersections resulting in indistinguishable field-features. Similarly in soccer, basketball and ice-hockey, parts of the field only contain a few circles or lines resulting in sparse field-features not sufficient for accurate homography estimation.

Sports	Field Size	Uniform Field Appearance	Camera Movement	Player Occlusion
Soccer	large	medium	medium	medium
American football	large	high	large	large
Ice hockey	medium	medium	large	medium
Basketball	medium	medium	medium	large
Tennis	small	small	small	small

Table 1. Factors that directly or indirectly affect how challenging the problem of field-registration for different sports is. Also listed is the degree to which these factors make the problem challenging.

b. Narrow field of view: this problem-source is common for sports such as American football where the camera needs to zoom in substantially to capture the action, which results in fields of view without sufficient distinct features.
c. Field occlusion: this problem-source arises when multiple players can group together in a way that occludes majority of the distinct key-features present on the field.

Key contributions: To address the challenge of insufficient distinct field-features for sports-field registration, we make the following two key contributions:

• First, unlike previous approaches that only use sparse locally-salient keypoints (*e.g.* corners and line intersections), we detect a grid of uniformly distributed keypoints over the entire field, thereby significantly increasing the likelihood that any camera pose over the duration of a gameplay would have sufficient number of field-keypoints visible in it. Additionally, we improve the detection of this uniform grid of keypoints by incorporating: (a) dilated convolution [7] and (b) non-local blocks [35] in our detection network. Our use of additional keypoints arranged as a uniform grid along with the use of RANSAC [12] substantially improves the robustness of our homography estimation.

• Second, for frames where our keypoints-based homography estimate is not well-conditioned, we propose the use of novel dense features to further optimize it. These dense features are defined as the normalized distance-map of each pixel in the field template to its nearest line and key region (*e.g.* yard-line numbers in American football). Although these dense features do not have correspondence information, they provide strong cues to optimize the keypoints-based homography by performing its local alignment. Additionally, we apply pixel-wise weight in our alignment-loss to ensure that the lines and regions that are visible in a frame are weighted more than the ones occluded by players. This adaptive weighting mechanism naturally provides robustness against field occlusions caused by the gathering of multiple players within a narrow field-of-view.

Since our keypoints and dense features are predicted using the same multi-task network, our method is able to run in real-time under HD resolution using commodity hardware. We present a comparative empirical analysis of multiple state-of-the-art field-registration approaches using a newly collected dataset with 192 video-clips from 5 different sports spanning a wide range of field appearance and camera poses, and demonstrate that our approach offers better accuracy compared to all of the considered approaches.

2. Related Work

Homography Estimation: The traditional framework for homography estimation is to first extract distinct keypoints using local feature detectors such as SIFT [24] and ORB [32], and then find correspondences by comparing feature descriptors, followed by estimating homography matrix using RANSAC [12] with DLT [15] or non-linear optimization [26]. This approach is limited by the accuracy of the local feature detector which is not robust to large appearance variations and non-static objects. In recent years, deep networks have been successfully used for detecting keypoints for human pose estimation [8, 4, 13, 1, 29, 11]. These networks have an encoder-decoder structure [30] and skip connections [31] which help the network to capture context from a larger image area than the local descriptors.

Another approach for homography estimation is to directly regress the homography parameters through a deep network. For instance, work in [10] utilizes deep network to regress the homography matrix which is re-parameterized as 4 control points. Work in [28] uses an unsupervised approach to train deep network for homography regression where the network predicts the homography to warp the image, and the photometric loss between two images is used to compute the gradient for back-propagation. Lastly, work in [22] uses a multi-scale network to estimate homography while pixels of moving objects are handled explicitly.

Sports-Field Registration: Sports-field registration is a special case of homography estimation as it focuses on sports-fields which have known fixed structure. Work in [17] formalizes the field registration problem using a Markov Random Field where the energy function is defined in terms of binary semantic cues such as surfaces, lines, and circles which are obtained from a deep network. Both [5] and [34] use synthetic data to generate the database which contains field template under different camera poses. They first segment the field surface and then use siamese network to retrieve the field template from the database. Work in [5] uses Lucas-Kanade algorithm to refine the homography and [34] uses differentiable warping for refinement. Work in [19] first uses a single network to directly regress the 4-point homography parameters as an initial estimate, and then another network to estimate the registration error. The initial estimation is optimized using the gradients provided by the registration-error network. Work in [9] uses keypoints from both field and players to estimate the homography. The keypoints are manually defined on the field and detected using a deep network. This method is difficult to generalize to

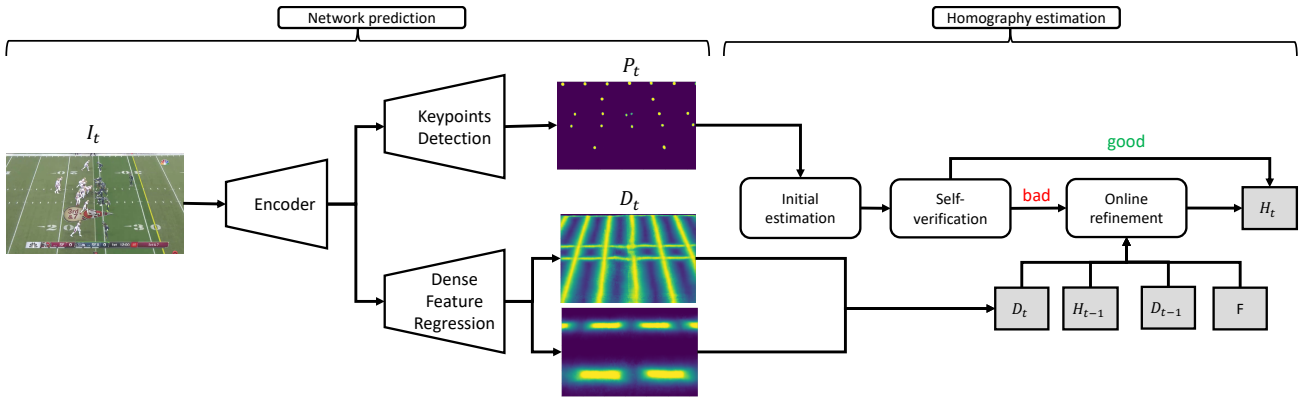


Figure 2. **Overview of the proposed framework** – given the current image $\mathbf{I}_t$, we detect its keypoints $\mathbf{P}_t$ and regress the two-channel dense features $\mathbf{D}_t$ using our multi-task network. $\mathbf{P}_t$ is used for the initial homography estimation which then goes through the self-verification step to compute its goodness score. If this goodness score is below a certain threshold, we further optimize the homography by aligning $\mathbf{D}_t$ with previous dense feature $\mathbf{D}_{t-1}$ and template feature map $\mathbf{F}$.

broadcast videos because the players' 3-D locations are assumed to be known. Work in [6] explicitly address the situation when the four-point pairwise correspondence is not satisfied due to the lack of feature points. It uses pan, tilt and zoom instead of homography to represent the transformations such that only two points are needed for estimation. The limitation for this approach is that it assumes that the camera's base-location and orientation are known *a priori*.

Key Differences of Our Approach: While [17, 5, 34] also use dense features for online homography optimization, our approach of incorporating dense features for warping based on pixel-wise weighted alignment loss is more efficient and robust to large camera movements. Also, unlike [9, 6] which require prior knowledge of players or camera location to address lack of distinctive features, we show that combining uniformly distributed grid of keypoints with dense features lets us overcome this challenge substantially without requiring any additional information about players or camera location. Lastly, in addition to using encoder-decoder network structure [4, 13, 1, 31], we also integrate the recently proposed non-local block layers [35] in our network to incorporate longer-range scene-context.

3. Method

As illustrated in Figure 2, our framework consists of two main components: (a) network prediction and (b) homography estimation. During network prediction, our multi-task deep network (see § 4.2 for details) takes in an image $\mathbf{I}_t$ as input and predicts its keypoints $\mathbf{P}_t$ and frame dense-features $\mathbf{D}_t$ (see § 3.2 for details). The homography estimation step then uses the feature map $\mathbf{F}$, as well as keypoints and homography related information from the current and previous frames to estimate the current homography $\mathbf{H}_t$. We now describe various parts of our framework in more detail.

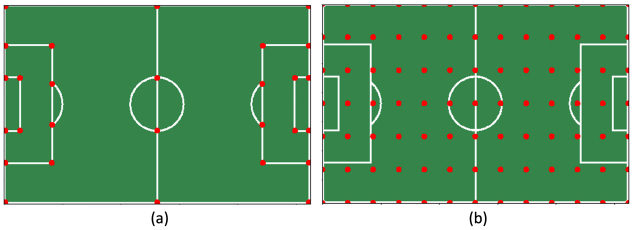


Figure 3. Comparison of keypoints. (a) The keypoints are defined on field corners or line intersections. (b) The keypoints are defined on a uniform grid across the entire field template.

3.1. Keypoints Detection

Instead of detecting sparse keypoints *e.g.* field corners and line intersections (see Figure 3-a), we detect a grid of keypoints distributed uniformly on the field template (see Figure 3-b). Sparse keypoints do not sufficiently cover various parts of the field, and therefore can lead to inaccurate homography estimation. In contrast, our use of uniform grid-based key-points significantly increases the likelihood that each part of the field is covered by well-conditioned subset of grid-keypoints, thereby increasing our ability to estimate homography accurately.

Accurate detection of uniformly distributed grid-based keypoints requires us to increase the receptive field of our network in order to capture longer range context-area around each pixel. We employ two important means to incorporate larger receptive field in our networks:

a. Dilated Convolutions: we use dilated convolutions [7] to increase the size of the convolution kernel at the last two blocks of the encoder network with a dilation size of 2.

b. Non-Local Blocks: in addition to the increase in the kernel size achieved via the use of dilated convolutions, we use non-local blocks [35] to capture global pixel-dependencies

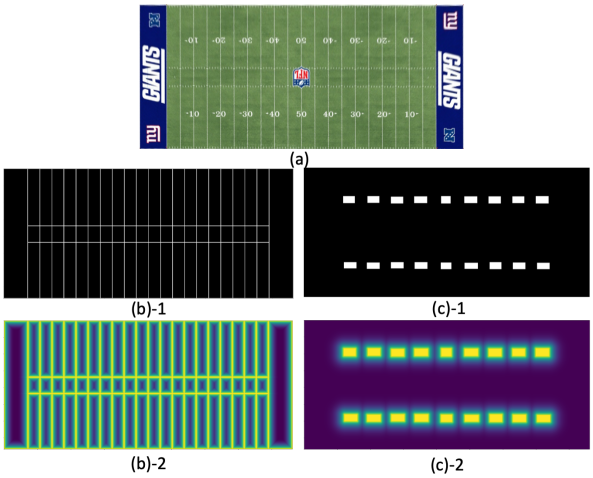


Figure 4. **Visualization of dense template-features –** (a) Template of American football-field. **(b)–1** Annotated binary line-map. **(b)–2** Line features defined as normalized distance between black pixels and their nearest white-pixels in the binary line-map. **(c)–1** Annotated binary region-map. **(c)–2** Region features defined as normalized distance between black pixels and their nearest white-pixels in the binary region-map.

over the entire frame. This non-local block is added after each of the last two residual blocks of our encoder network.

During training, keypoints from the field template are projected to each frame using ground truth homography to generate the ground truth positions and labels of keypoints. The output from keypoint detection head is N+1 heatmaps where N channels for keypoints and 1 channel for background. To deal with the unbalanced negative and positive data, we dilate each keypoint by 10 pixels and use weighted cross-entropy loss. The weight for each keypoint is set to 100 while for non-keypoint pixels it is set to 1. During inference, the heatmaps are decoded into the keypoint set $\mathbf{P}_t = [(x_1, y_1, c_1), (x_2, y_2, c_2), ..., (x_{\mathrm{N}}, y_{\mathrm{N}}, c_{\mathrm{N}})]$ where (x, y) are the coordinates and c is the label. These keypoints and their corresponding field points are then used to estimate the homography parameters using the DLT [15] algorithm.

3.2. Dense Feature Regression

Although the coverage of our grid based keypoints is significantly better than local salient keypoints, there can still be frames that do not have enough grid-keypoints to use DLT [15] algorithm due to small field-of-view or heavy field occlusion. For frames where this issue happens, we use our network to regress the dense frame-features directly and align them with dense template-features in order to further refine the initial key-point based homography estimate.

To this end, we introduce two types of dense template-features: (a) line feature and (b) region feature (see Figure 4 for an example using American football field). These dense features are defined as the normalized distance maps of pixels to their nearest lines and regions respectively. We annotate the lines and regions from the field template once (Figure 4 (b)-1 and (c)-1), and convert them to continuous distance maps by a distance transform and normalize them from 0 to 1 to form a probability map (Figure 4 (b)-2 and (c)-2). During training, the dense feature maps are back-projected to each frame using ground truth homography. Line features and region features are predicted from two separate heads of the network and concatenated as a two channel dense frame-feature $\mathbf{D}_t$. We use $\mathbf{L}_1$ loss between $\mathbf{D}_t$ and the warped feature during network training.

3.3. Homography Estimation

We now discuss how to use the keypoints $\mathbf{P}_t$ and dense features $\mathbf{D}_t$ to estimate the homography $\mathbf{H}_t$. The keypoints $\mathbf{P}_t$ are used to estimate the initial homography and also the inlier point-set. This is followed by a self-verification step that decides if the online optimization step is needed. In case the online optimization is not required, the initial homography estimate is used as the final homography $\mathbf{H}_t$, otherwise we compute $\mathbf{H}_t$ using $\mathbf{F}$, $\mathbf{D}_t$, $\mathbf{D}_{t-1}$ and $\mathbf{H}_{t-1}$ by optimizing the alignment loss.

3.3.1 Initial Homography Estimation

Given detected keypoints $\mathbf{P}_t$ from frame $\mathbf{I}_t$, we directly estimate the initial homography $\mathbf{H}_t^{\text{initial}}$ using DLT [15] and RANSAC [12]. The inlier point-set $\mathbf{P}_t^{\text{inliers}}$ is also identified by $\mathbf{H}_t^{\text{initial}}$. We set $\mathbf{P}_t^{\text{inliers}}$ to an empty-set if the estimated $\mathbf{H}_t^{\text{initial}}$ is degenerate due to poor configuration of $\mathbf{P}_t$.

3.3.2 Self-Verification

The self-verification step checks the goodness of the initial homography estimate $\mathbf{H}_t^{\text{initial}}$ and decides if online optimization is needed to further improve it. To verify the goodness of $\mathbf{H}_t^{\text{initial}}$, we check two conditions:

(a) Goodness of $\mathbf{P}_t^{\text{inliers}}$: the goodness of $\mathbf{P}_t^{\text{inliers}}$ is decided by their position configuration. More inlier points with more uniform position distribution result in a higher goodness score. Similar to [33], we divide the image into grids at three resolutions: $2 \times 2, 4 \times 4, 8 \times 8$ and compute the goodness scores at each resolution as the number of cells occupied by the points.

(b) Consistency between $\mathbf{H}_t^{\text{initial}}$ and $\mathbf{H}_{t-1}$: to verify the temporal consistency of homography estimation, we compute the intersection-over-union (IOU) measure between the binary maps projected using $\mathbf{H}_t^{\text{initial}}$ and $\mathbf{H}_{t-1}$.

We use $\mathbf{H}_t^{\text{initial}}$ as final $\mathbf{H}_t$ if the goodness of $\mathbf{P}_t^{\text{inliers}}$ and the consistency of $\mathbf{H}_t^{\text{initial}}$ are above certain thresholds. Otherwise, we perform online homography refinement. The thresholds for the two conditions are chosen by cross-validation on training data.

3.3.3 Online Homography Refinement

Our objective function of online homography refinement optimization is listed in Equation 1.

$$\begin{aligned} \mathbf{H}_t &= \arg\min_{\mathbf{H}} \lambda_f \mathcal{L}_f + \lambda_s \mathcal{L}_s \\ \mathcal{L}_f &= \frac{||(\mathbf{D}_t - \text{warp}(\mathbf{H}, \mathbf{F})) \cdot \mathbf{D}_t||_1}{||\mathbf{D}_t||_1} \\ \mathcal{L}_s &= ||\mathbf{D_t} - \text{warp}(\mathbf{H}\mathbf{H}_{t-1}^{-1}, \mathbf{D}_{t-1})||_1 \end{aligned} \tag{1}$$

The function $\text{warp}(\mathbf{H}, \mathbf{F}))$ warps feature map $\mathbf{F}$ using homography $\mathbf{H}$ and *differentiable* bilinear sampling [18]. Here, operation $\cdot$ is element-wise multiplication and $||\quad||_1$ is vectorized $\mathbf{L_1}$ loss.

Intuitively, we want to encourage homography to have good alignment between predicted dense frame-feature $\mathbf{D}_t$ and dense template-features $\mathbf{F}$. Additionally, to incorporate temporal smoothness, we want the current dense frame-feature $\mathbf{D}_t$ to have good alignment with the previous dense frame-feature $\mathbf{D}_{t-1}$. To achieve this, we define two losses:

(a) Matching Loss ($\mathcal{L}_f$) – measures pixel-wise weighted difference between the current dense frame-feature $\mathbf{D}_t$ and the warped field feature using the current homography estimate $\mathbf{H}$ and the dense template-features $\mathbf{F}$. We directly use $\mathbf{D}_t$ as pixel-wise weights so that the mismatched pixels closer to lines and key regions contribute more to the loss.

(b) Tracking Loss ($\mathcal{L}_s$) – computes difference between $\mathbf{D}_t$ and warped $\mathbf{D}_{t-1}$ using relative homography $\mathbf{H}\mathbf{H}_{t-1}^{-1}$. It encourages temporal consistency between neighboring estimations.

This optimization process is only used during inference, and done by back-propagation with Adam optimizer [21]. The loss weights λ_f and λ_s are set to 0.9 and 0.1 respectively.

4. Experiments

4.1. Datasets

To the best of our knowledge, the only publicly available dataset with homography annotation is the Soccer World Cup dataset [17] which has 209 images for training and 186 images for testing. In order to do a more comprehensive evaluation and facilitate further research in sports-field registration, we collected a new dataset called **SportsFields** from five different sports including soccer, American football, ice hockey, basketball and tennis. For each sport, we manually selected game-play video segments which cover a large variation of field appearance and environmental conditions (*e.g.* rain, snow, harsh sunlight, *etc.*). The training and testing images are sampled for manual annotation at 1 FPS while ensuring that images from the same videos do not get included in both train and test splits. For testing, we run our method on each frame of the test videos and compute the evaluation metrics only for frames for which we have annotations available. Various statistics for **SportsFields** dataset is shown at Table. 2.

Sport	Number of videos	Number of training images	Number of testing images
Soccer	60	396	321
American football	30	502	333
Ice hockey	23	321	199
Basketball	21	359	130
Tennis	27	255	151

Table 2. **SportsFields data statistics**: number of videos, and train and test images for each sport in SportsFields dataset.

4.2. Implementation Details

a– Network: Our multi-task network has an encoder-decoder structure [30] based on ResNet-18 [16]. The decoder is composed of four deconvolution layers to output the final feature map which has $1/4$ resolution of the input image. Similar to U-Net [31], we add skip connections between encoder and decoder to fuse the features at different scales. Additionally, dilated convolution and Non-Local block [35] are added in the last two residual blocks of encoder. Three heads are attached on top of the last feature map for keypoint detection, line feature and region feature regression respectively. We use L1 loss for line feature and region feature regression, and cross-entropy loss for keypoint detection with weights of each loss set to 1.

b– Hyperparameters: We train separate networks for different sports with the same hyperparameters. All networks are trained for 200 epochs using learning rate of $1e-4$ and then another 100 epochs using learning rate of $1e-5$. We use Adam optimizer [21] with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for all epochs. For online optimization, we run Adam optimizer with learning rate 0.001 for 50 iterations. The weights λ_s and λ_f are set to 0.1 and 0.9 respectively. The re-projection threshold for RANSAC [12] is 10. For self-verification, the goodness threshold for keypoints configuration is set to 100 while the IOU threshold for homography consistency is set to 0.6.

4.3. Evaluation Metrics

We report four metrics for evaluation purposes including two types of Intersection Over Union (IOU), as well as the projection error and the re-projection error. For each metric, we report both mean and median values.

a– IOU: There are two types of IOU metrics commonly used by previous approaches. The first type, *i.e.* IOU_{part}, is computed between binary masks produced by projecting the field-mask in a video-frame while using predicted homography and ground-truth homography. The second type, *i.e.* $\text{IOU}_{\text{whole}}$, is computed between field-template binary mask

and its projection using product of predicted homography-inverse and ground truth homography (see [9] Appendix C).

Work in [9] pointed out that $\text{IOU}_{\text{whole}}$ is better because it also compares the invisible field area. However, the registration error of the invisible field cannot be measured because the ground truth homography is obtained using the visible field area and guarantees the projection correctness of the visible area only. Therefore, we prefer using IOU_{part} in conjunction with $\text{IOU}_{\text{whole}}$. Both types of IOU metrics have the obvious flaw that they ignore the dense correspondences of points, and can have large value even when the predicted homography is completely wrong, *e.g.* the flipped field has the same IOU as the non-flipped one. To overcome these limitations of IOU metric, we also compute two additional evaluation metrics that are described below.

b– Projection Error: The projection error is measured by the average distance in actual scale (meters) between projected points using predicted homography and the points using ground truth homography. This metric is invariant to different image resolutions. We uniformly sample 2500 pixels from visible field area of the camera image and project them to the field to compute the distance. The actual field dimensions we use for soccer, American football, basketball, ice hockey and tennis are (in meters) 100×60, 110×49, 29×15, 61×26 and 24×11 respectively.

c– Re-projection Error: We follow [9] in using the re-projection error as one of our evaluation metrics. Recall that re-projection error is defined as the average distance between points re-projected in the video-frame using predicted homography and those re-projected using ground-truth homography. We compute the pairwise distances between the re-projected points and normalize them by image height to compute the final average re-projection error.

4.4. Baselines

We compare our method against other state of the art approaches [20, 3, 5, 19, 34, 9] on the public Soccer World Cup dataset [17]. We use the inference code and pre-trained model released by [19] to compute all metrics. We use the results of approach [5] reported in [19] and the results of [20] reported in [9]. To have a fair comparison with [9] we take the results of approach ours-w/o-players reported in their paper for IOU and re-projection error. We implement [3] following the details given in the paper.

For the sake of completeness, we also derive three variants of our approach (listed below) and present their results. The first two variants are single frame based methods and therefore do not need consecutive frames for testing.

• **ours – keypoints only**: we use only detected keypoints to compute the homography. Online optimization is not used.

• **ours – alignment**: we use detected keypoints to compute the initial homography and do online optimization with only matching loss so the dense features and homography from previous frame are not used.

• **ours – tracking**: we use detected keypoints to compute initial homography and do online optimization with only tracking loss so the good homography estimates are tracked.

• **ours – all**: our complete approach with all components.

Method	fps	$\text{IOU}_{\text{part}}(\%)$		$\text{IOU}_{\text{whole}}(\%)$		Proj. (meter)		Re-Proj.	
	mean	mean	median	mean	median	mean	median	mean	median
Posenet [20]	19	78.5	82.9	63.4	66.3	5.22	4.62	0.243	0.163
Eric et al. [3]	5	93.4	94.5	86.4	88.8	1.26	1.01	0.037	0.031
Chen et al. [5]	-	94.7	96.2	89.2	91.0	-	-	-	-
Wei et al. [19]	0.6	95.1	96.7	89.8	92.9	1.21	0.74	**0.017**	**0.012**
Long et al. [34]	**250**	93.2	96.1	88.3	92.1	-	-	-	-
Leonardo et al. [9]	9	-	-	90.5	91.8	-	-	0.018	0.012
ours – keypoints Only	50	95.8	**97.2**	91.5	93.3	**0.82**	**0.61**	0.019	0.015
ours – alignment	2	**95.9**	97.1	**91.6**	**93.4**	0.84	0.65	0.019	0.014

Table 3. Quantitative results of homography estimation on Soccer World Cup dataset.

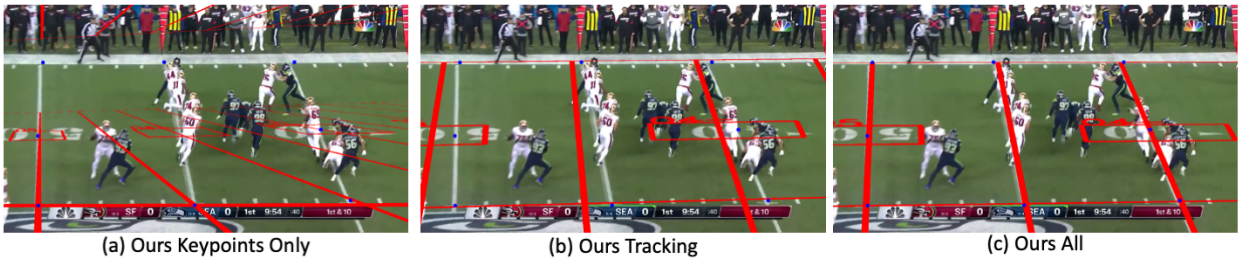
Method	$\text{IOU}_{\text{part}}(\%)$		$\text{IOU}_{\text{whole}}(\%)$		Proj. (meter)		Re-proj.	
	mean	median	mean	median	mean	median	mean	median
Posenet [20]	13.2	12.2	0.6	0.3	39.23	33.65	1.210	0.954
Eric et al. [3]	38.3	41.1	38.8	39.5	9.64	5.74	5.096	0.702
Wei et al. [19]	18.1	17.4	5.8	5.1	28.55	25.81	1.032	0.876
ours – keypoints only	66.7	97.1	64.3	95.1	50.60	0.21	1.832	0.011
ours – alignment	67.1	97.1	66.5	95.2	40.21	0.20	1.873	0.011
ours – tracking	81.7	97.3	82.1	95.7	1.83	0.21	0.192	0.012
ours – all	**96.2**	**97.4**	**94.2**	**96.1**	**0.22**	**0.18**	**0.015**	**0.011**

Table 4. Quantitative results of homography estimation on American football dataset.

4.5. Results

a– Soccer World Cup: Table 3 presents the evaluation of our approach on the publicly available Soccer World Cup dataset [17]. Like previous approaches, our model is trained using 209 images and tested on 186 images. We only evaluate our single frame based method variants **ours – keypoints only** and **ours – alignment** because the testing data are not consecutive frames. The FPS of inference is tested using GTX 1080ti with CUDA 9.0. From Table 3 we can see that our method variant with improved keypoints definition and detection not only achieves better IOU and projection error, but also runs faster than most of other approaches. The benefit of using online optimization with dense features is minor in this data as the results using keypoints alone are already quite accurate.

b– American Football: We compare the various variants of our approach with [20, 3, 19] on American football part of our SportsFields data as it highlights more challenges of the problem at hand than the other four sports do. The quantitative results are reported in Table 4. We can see that the direct regression approaches [20, 19] do not work well on this data, and the approach in [3] is better than [20] and [19] but far from the accuracy it achieves on Soccer World Cup dataset. The likely reasons for this difference include: (i) the variation of homography on American football is much larger than other sports because of the large field size and



(a) Ours Keypoints Only (b) Ours Tracking (c) Ours All

Figure 5. Qualitative comparison of the results for different variants of our approach. (a) The homography is estimated only by the detected keypoints from current frame. (b) The homography is optimized by using tracking loss which aligns the current dense features with previous dense features. However, it can be seen that the homography starts to drift because of the heavy field occlusions caused by the presence of multiple players in a small field-of-view. (c) The homography is optimized by using both tracking loss and matching loss.

Figure 6. Qualitative results of our approach (**ours – all**) for challenging examples from five different sports in our SportsFields data. Four examples of each sport are visualized with sport field lines (red) and keypoints (blue) overlayed on the image. The last column shows one failure example for each sport. More qualitative results are presented in supplementary materials.

small camera field-of-view. The network therefore can have difficulty to learn large homography variations, and (ii) the field appearance is quite uniform and the distinguishable features are sparse. The network has heavy down-sampling at later stages and therefore cannot capture the distinguishable features well.

Our method variant (**ours – keypoints only**) that uses only keypoints is able to achieve significantly better accuracy than existing approaches. The large difference between the median and mean for each metric indicates that many frames get accurate estimates while the estimates for some frames are completely wrong as shown in Figure 5 (a). Our

Method	fps	$IOU_{part}(\%)$		$IOU_{whole}(\%)$		Proj. (meter)		Re-Proj.	
	mean	mean	median	mean	median	mean	median	mean	median
Soccer	35	97.5	97.9	92.9	93.5	0.33	0.12	0.014	0.011
American Football	25	96.2	97.4	94.2	96.1	0.22	0.18	0.015	0.011
Ice Hockey	35	95.2	96.1	91.9	93.0	0.35	0.30	0.023	0.021
Basketball	35	97.4	97.7	92.3	93.9	0.19	0.17	0.013	0.010
Tennis	45	99.9	99.9	97.5	97.7	0.09	0.08	0.009	0.009

Table 5. Quantitative results of homography estimation of our approach *ours – all* on SportsFields dataset.

method variant (**ours – tracking**) that uses tracking loss for alignment improves the keypoints only variant by tracking the good estimates from previous frames, but it starts to drift due to the heavy field occlusion as shown in Figure 5 (b). With both tracking loss and matching loss, our final approach (**ours – all**) further improves the results (shown in Figure 5 (c)) by aligning the dense features with both previous frame and field template.

c– SportsFields: Finally, we evaluate our full approach (**ours – all**) on all five sports using our SportsFields dataset. The quantitative results are reported in Table 5. For all sports the IOU_{part} is above 95% and the projection error is within half meter which demonstrates the good generalization ability of our approach. Among the five sports, tennis gets the best registration accuracy because of the limited camera movement and the fully visible field observed in Tennis for most of the times. The inference FPS of our approach is mainly affected by the percentage of frames which need online optimization which accounts for 90% of the total computation time. Processing American football data takes longest time because the small field-of-view and uniform field appearance observed in American football leads to more frequent online optimization than other sports. Five qualitative examples for each sport are provided in Figure 6. The results demonstrate that our approach handles various challenging situations well, *e.g.*, the middle field of soccer and basketball court, and small field-of-view of American football. The last column of Figure 6 shows failure-cases mainly caused by: (a) online optimization failing to align small field areas, (b) heavy motion blur, and (c) unexpected camera occlusions (*e.g.* audience raising hands).

4.6. Ablation Study

a– Keypoints Definition: To verify the benefit of our uniformly defined grid-keypoints, we compare the homography estimation between the two types of keypoints (Figure 3) in Table 6 on World Cup dataset. The uniform keypoints clearly perform significantly better than the locally defined keypoints since the corners and intersections cannot cover the full field, and therefore the limited number of keypoints and their poor positional configuration can often lead to inaccurate homography estimation.

b– Grid-Keypoints Detection: We evaluate the accuracy of grid-keypoints detection task using our SportsFields dataset. We also evaluate the benefit of dilated convolution and Non-Local block in this regard. The average precision and recall of all keypoints are computed in Table 7. The accuracy of keypoints detection indicates the difficulty of using only keypoints for field registration. Among five sports, American football has the worst accuracy due to the uniform field appearance and the small camera field-of-view. Tennis gets almost perfect result because of the large field-of-view and small camera movement. With dilated convolution and non-local block, four of the more challenging sports (all but tennis) in SportsFields dataset get a significant improvement which demonstrates that using large receptive field and long-range context are critical for uniform grid-keypoints detection.

Method	$IOU_{part}(\%)$		$IOU_{whole}(\%)$		Proj. (meter)		Re-Proj.	
	mean	median	mean	median	mean	median	mean	median
grid keypoints	**95.8**	**97.2**	**91.6**	**93.3**	**0.82**	**0.61**	**0.019**	**0.015**
local keypoints	90.6	95.3	81.9	88.9	3.51	0.92	0.164	0.024

Table 6. Comparison of different types of keypoints on homography estimation on Soccer World Cup dataset.

Sport type	#Keypoints	Without dilation+NL		With dilation+NL	
		Precision	Recall	Precision	Recall
soccer	91	0.66	0.66	**0.75**	**0.76**
American football	168	0.39	0.36	**0.55**	**0.49**
ice hockey	75	0.45	0.43	**0.53**	**0.55**
basketball	50	0.71	0.68	**0.75**	**0.72**
tennis	15	0.97	0.99	**0.99**	**0.99**

Table 7. Comparison of keypoints detection on different network structures. Here NL implies non-local block [35].

5. Conclusions and Future Work

We presented a general framework to register sports-fields as they appear in broadcast sports videos particularly when there are not enough distinguishable features available on the field. The key technical contributions we made to overcome this challenge are: (i) detecting a uniform grid of keypoints that cover the entire field, and (ii) using dense field-features to further refine the keypoint based homography estimate. We collected a new and significantly more comprehensive field registration dataset (called SportsFields) which covers five sports with large variations of field appearance and camera pose. We demonstrated that our framework is able to register sports-fields more accurately than previous approaches under challenging capturing conditions and runs in real-time for HD resolution.

Going forward, we will explore combining bundle adjustment [15] with dense features to jointly optimize homograpy estimate for all video-frames instead of frame-by-frame processing. Another direction we want to explore is using semi-supervised learning to minimize labelling effort.

References

[1] Newell A, Yang K, and Deng J. Stacked hourglass networks for human pose estimation. 2016.

[2] N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 1992.

[3] Eric Branchmann and Carsten Rother. Learning less is more - 6d camera localization via 3d surface regression. 2018.

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. 2017.

[5] J. Chen and J. J. Little. Sports camera calibration via synthetic data. 2019.

[6] Jianhui Chen, Fangrui Zhu, and James J. Little. A two-point method for ptz camera calibration in sports. 2018.

[7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *ArXiv*, abs/1706.05587, 2017.

[8] Xianjie Chen and Alan L Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. 2014.

[9] Leonardo Citraro, Pablo Márquez-Neila, Stefano Savarè, 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.

[10] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Deep image homography estimation. 2016.

[11] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. 2018.

[12] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. 1981.

[13] Rıza Alp Güler, George Trigeorgis, Epameinondas Antonakos, Patrick Snape, Stefanos Zafeiriou, and Iasonas Kokkinos. Densereg: Fully convolutional dense shape regression in-the-wild. 2017.

[14] Chris Harris and Mike Stephens. A combined corner and edge detector. *Proceedings of the 4th Alvey Vision Conference*, 1988.

[15] R. I. Hartley and A. Zisserman. Multiple view geometry. *Computer Vision. Cambridge University Press*, 2004.

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[17] Namdar Homayounfar, Sanja Fidler, and Raquel Urtasun. Sports field localization via deep structured models. *CVPR*, 2017.

[18] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial transformer networks. pages 2017–2025, 2015.

[19] 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. *WACV*, 2019.

[20] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. 2015.

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2014.

[22] Hoang Le, Feng Liu, Shu Zhang, and Aseem Agarwala. Deep homography estimation for dynamic scenes. 2020.

[23] V. Lepetit, F. Moreno-Noguer, and P. Fua. Epnp: An accurate o(n) solution to the pnp problem. 2009.

[24] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, pages 91–110, 2004.

[25] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004.

[26] Jorge J. Moré. The levenberg-marquardt algorithm: Implementation and theory. pages 105–116, 1978.

[27] Marius Muja and David G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. *International Conference on Computer Vision Theory and Applications*, 2009.

[28] T Nguyen, SW Chen, SS Shivakumar, CJ Taylor, and V Kumar. Unsupervised deep homography: A fast and robust homography estimation model. pages 2346–2353, 2018.

[29] Bruce Xiaohan Nie, Ping Wei, and Song-Chun Zhu. Monocular 3d human pose estimation by predicting depth on joints. 2017.

[30] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. 2015.

[31] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 9351:234–241, 2015.

[32] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. *International Conference on Computer Vision (ICCV)*, 2011.

[33] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. 2016.

[34] Long Sha, Jennifer Hobbs, Panna Felsen, Xinyu Wei, Patrick Lucey, and Sujoy Ganguly. End-to-end camera calibration for broadcast videos. 2020.

[35] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. 2018.