https://doi.org/10.31891/2307-5732-2025-359-63
УДК 004.93

SOROKIVSKYI OLEKSANDR
Ternopil Ivan Puluj National Technical University
https://orcid.org/0009-0006-6477-5878
e-mail: sasha.sorokivski@gmail.com
VOLODYMYR HOTOVYCH
Ternopil Ivan Puluj National Technical University
https://orcid.org/0000-0003-2143-6818
e-mail: gotovych@gmail.com

DESIGNING A NEURAL NETWORK ARCHITECTURE TO ACCELERATE CAMERA CALIBRATION IN SOCCER MATCH ANALYTICS

The article proposes a method to accelerate the camera calibration process in soccer match analytics while maintaining acceptable accuracy. The study focuses on modifying the High-Resolution Network (HRNet) architecture to reduce computational costs, making it suitable for real-time applications under limited hardware conditions. HRNet is a deep learning architecture known for maintaining high-resolution feature representations throughout its layers, which is particularly valuable for dense prediction tasks like keypoint detection and semantic segmentation. Unlike traditional models that downsample spatial information early, HRNet preserves detailed spatial features by processing multiple resolutions in parallel and continuously exchanging information across them.

The proposed by authors approach builds on HRNet's strengths while improving its efficiency through applying three key strategies: simplifying network structures, applying knowledge distillation to transfer information from larger models to smaller ones, and adopting multi-task learning to handle keypoint and line detection within a single unified model.

The study evaluates several HRNet variants, including the standard simplified versions (W32, W18), as well as new architectures developed by the authors: an ultra-compact version (W6) and a multi-task model. The analysis focuses on the trade-off between speed and accuracy.. These models are trained and tested on the SoccerNet 2023 dataset, which offers a large and diverse set of annotated soccer images from multiple camera viewpoints. The evaluation uses practical metrics that reflect both calibration accuracy and completeness across varied match conditions.

Obtained results show that the developed W6 model achieves up to a 270% increase in processing speed compared to the original HRNet, with only a moderate drop in performance (12%). Meanwhile, the proposed multi-task architecture delivers the highest accuracy among the larger models trained for the same number of epochs, while also improving processing speed even in the smaller variants. Based on the obtained results conclusion is made that these compact and multi-task architectures offer a practical solution for fast, automated camera calibration in real-world sports analytics.

***Keywords*: :** *Camera calibration, Soccer analytics, Deep learning, HRNet, Knowledge distillation, Multi-task learning.*

СОРОКІВСЬКИЙ ОЛЕКСАНДР, ГОТОВИЧ ВОЛОДИМИР
Тернопільський Національний Технічний Університет

РОЗРОБКА АРХІТЕКТУРИ ШТУЧНОЇ НЕЙРОННОЇ МЕРЕЖІ ДЛЯ ЗАДАЧІ ПРИШВИДШЕННЯ ПРОЦЕСУ КАЛІБРУВАННЯ КАМЕРИ В АНАЛІТИЦІ ФУТБОЛЬНИХ МАТЧІВ

У статті запропоновано метод прискорення процесу калібрування камери в аналітиці футбольних матчів при збереженні прийнятної точності. Дослідження зосереджене на модифікації архітектури High-Resolution Network (HRNet) ***з метою зниження необхідних для неї*** *обчислювальних витрат, що робить її придатною для застосування в реальному часі в умовах обмежених апаратних ресурсів. HRNet – це модель глибокого навчання, яка зберігає чіткі зображення на всіх етапах і добре підходить для задач, де важливо точно визначати деталі, наприклад, ключові точки або сегменти на зображенні. На відміну від традиційних моделей, які на ранніх етапах зменшують просторову роздільність, HRNet зберігає детальні просторові ознаки, обробляючи кілька різних маштабів паралельно та постійно обмінюючись інформацією між ними.*

Запропонований авторами підхід базується на перевагах HRNet і покращує її ефективність за допомогою застосування трьох ключових стратегій: спрощення структури мережі, застосування дистиляції знань для передачі інформації від більших моделей до компактніших, а також використання багатозадачного навчання для одночасного вирішення задач виявлення ключових точок і ліній у межах єдиної уніфікованої моделі.

У дослідженні оцінюються кілька варіантів HRNet, зокрема стандартні спрощені версії (W32, W18), а також нові архітектури, розроблені авторами: надкомпактна (W6) і багатозадачна модель. Аналіз проводиться з урахуванням балансу між швидкістю та точністю. Ці моделі навчаються та тестуються на наборі даних SoccerNet 2023, який містить велику та різноманітну колекцію анотованих зображень футбольних матчів з різних ракурсів. Для оцінки використовуються практичні метрики, що відображають як точність калібрування, так і його повноту за різних умов проведення матчів.

*Отримані результати показують, що розроблена модель W6 забезпечує до 270% приросту швидкості обробки порівняно з оригінальною HRNet при лише помірному зниженні точності (12%). Водночас запропонована багатозадачна архітектура демонструє найвищу точність серед більших моделей***,** *які навчалися протягом такої ж кількості епох, а також володіє підвищеною швидкістю обробки навіть за умови застосування її компактних варіантів***.** *На основі отриманих результатів робиться висновок, що ці компактні та багатозадачні архітектури є практичним рішенням для швидкого автоматизованого калібрування камер.*

Ключові слова*: калібрування камери, футбольна аналітика, глибоке навчання, HRNet, дистиляція знань, багатозадачне навчання.*

Стаття надійшла до редакції / Received 01.09.2025
Прийнята до друку / Accepted 15.11.2025

Introduction

The use of information technologies for addressing analytics tasks in the analysis of soccer matches makes it possible to substantially simplify and accelerate the analytical process itself, while also improving sporting outcomes

for individual players and teams. Thanks to the introduction of advanced technologies – particularly systems for tracking the movements of players and the ball – it has become possible to generate datasets whose analysis yields valuable insights into the dynamics of sporting competitions. Such technologies have fundamentally changed approaches to the analysis of sporting events, enabling coaches to make more informed decisions [1, 2].

Today, deep learning and computer vision methods are widely used in sports analytics. With modern methods and algorithms, researchers and sports managers can extract valuable information about gameplay directly from video footage, without the need for invasive sensors attached to athletes' bodies [3]. This approach offers several advantages, including the ability to collect data on multiple athletes simultaneously and the potential for retrospective analysis of matches that have already been played.

Camera calibration is a crucial stage in many computer vision tasks. It is the process of determining both the internal and external parameters of a camera, which is necessary for accurately transforming 2D image data into informative 3D representations of the sporting environment (the pitch) [4, 5]. In the context of soccer matches, camera calibration enables the precise projection of a player from a video frame or photograph onto their actual location on the field. This is important both for the individualized analysis of a player's actions and for decision-making tasks such as determining offside or verifying whether a goal has been scored [4, 6, 7, 8].

When solving the camera calibration problem, speed is critically important and depends primarily on available computational resources. The ability to deliver rapid, near-real-time data processing is decisive for providing actionable insights and supporting decision-making during matches and training sessions [9, 10, 11, 12], especially for sports clubs and organizations with limited budgets – and thus limited computing power [13, 14].

This article proposes an approach that reduces the time required for camera calibration with varying degrees of performance trade-offs. Our approach employs reduced and modified versions of well-known artificial neural network architectures, including the use of knowledge distillation. We present several architectural variants that require training only a single model to achieve competitive results.

Related works

Overview of existing solutions

Historically, early solutions to the camera calibration problem relied primarily on classical computer vision algorithms, particularly those that extracted low-level features from images. While these approaches provided a foundational basis, their effectiveness in the field is limited by external factors – uneven illumination, shadows, and occlusion artifacts typical of sports broadcasts. In particular, the Scale-Invariant Feature Transform (SIFT) algorithm [15] demonstrated strong robustness in detecting and matching keypoints under changes in camera viewpoint, which proved useful, for example, for identifying intersections of field markings [16]. The Hough transform has also been widely used, notably for detecting straight lines in image frames, enabling automated localization of soccer field lines [17, 5]. However, the dependence on low-level visual features limits the adaptability of these methods to complex conditions, motivating a shift toward more robust and semantically rich image-analysis models [18, 19, 5].

With the development of deep learning, a new era of camera calibration methods emerged, characterized by more robust feature extraction and substantially higher accuracy. For example, [19] proposed a semantic segmentation approach that assigns each image pixel to one of six categories: vertical lines, horizontal lines, side circles, the center circle, grass, and stands with spectators. Such fine-grained pixel labeling provides a deeper and more coherent understanding of the soccer field's spatial layout compared with classical line and ellipse detection techniques. In [20], traditional geometric annotation was extended by using the image centroids of players within the frame as control points for calibration. Although this idea can improve the accuracy of perspective estimation, its practical implementation requires careful alignment among multiple cameras to correctly project detected local player centroids onto the field plane, complicating deployment when the number of cameras is limited.

Contemporary studies calibrating moving cameras employ deep neural networks (DNNs) capable of directly estimating projection parameters. In [21], a DNN was trained to regress a homography parameterization directly from a single input frame, thereby obtaining both intrinsic and extrinsic camera parameters within a single computational module with minimal preprocessing. The study [22] introduced the multitask SFLNet architecture, in which a single convolutional regressor predicts a metric field model as an eight-dimensional vector, performs semantic segmentation of the frame into field, player, and background regions, and constructs an adjacency matrix of labels to encode the mutual spatial arrangement of key scene points. Integrating these components in a single model yielded a substantial increase in the accuracy of projection-parameter estimation.

In established approache such as TVCalib [23] and PnLCalib [24] – accuracy and reliability are typically improved through iterative or multi-stage optimization, intensive keypoint and line detection, and extensive search in geometric parameter space. Execution time, however, is rarely treated as a critical metric.

The approach proposed in this work achieves high processing speed while maintaining competitive accuracy, thereby closing the gap created by methods focused exclusively on maximizing accuracy.

Established Approaches to the Camera Calibration Problem

This work builds on the methodology described in Falaleev et al. [25], Guti´errez P´erez et al. [26], and Guti´errez P´erez and Agudo [24]. In the first stage, the soccer pitch is modelled using lines, circles, and semicircles that reproduce the field markings and goal structure. Based on the known real-world positions of these lines, keypoints are generated sequentially: first line–line intersections, then extended intersections and points of tangency to ellipses, as well as additional control points to ensure a complete grid.

Keypoint and line detection is performed using one or two encoder–decoder convolutional neural networks. For each predefined point and line, heatmaps with Gaussian peaks are produced to indicate their locations, and an additional contour channel is introduced to sharpen edges and silhouettes.

The camera parameters (intrinsic and extrinsic) are estimated using the standard full-perspective projection model, where initial values are computed analytically without iterations, and subsequent nonlinear refinement is carried out by maximizing a likelihood function based on correspondences between the model's 3D field coordinates and their 2D projections. The plane-to-image homography is extracted from the camera projection matrix using DLT and RANSAC applied to a subset of keypoints. As a result, a single pass over a full image frame simultaneously generates predictions for the locations of keypoints and lines. Final calibration parameters are determined via a heuristic voting mechanism that favors solutions with the smallest reprojection error.

In the original study, an integrated refinement module for points and lines was included to enhance calibration accuracy and robustness. Within this module, information about lines and keypoints was used jointly, and camera parameters were obtained by minimizing an appropriate error function in position–orientation space. In the present work, this module is omitted due to its substantial impact on the algorithm's runtime.

Results

Proposed Approach to Accelerating the Calibration Process

This work proposes an approach aimed at increasing the runtime efficiency of an artificial neural network model while preserving the core principles of the original method.

Within the proposed framework, the High-Resolution Network (HRNet) is employed to compute the homography matrix. HRNet is optimized by downsizing the model, introducing knowledge distillation, and adopting a multitask architecture.

We analysed reduced variants of HRNet – particularly the "Small HRNet" – by decreasing the network's depth and width. This lowered the parameter count and computational demand [27]. In addition, we designed an even smaller HRNet variant, which further accelerated inference.

During knowledge distillation, a compact student model was trained to imitate the behaviour of a teacher model, enabling network compression with minimal loss in accuracy – an established practice for deploying deep models under constrained compute [28]. Transferring knowledge from the original HRNet to the reduced version preserved accuracy while increasing algorithmic speed.

To enable solving multiple tasks within a single architecture, we employ a multi-head network in which two HRNets with distinct output heads are unified into a single structure with a shared backbone. This design simultaneously predicts the locations of keypoints and lines. Multitask training fosters richer feature representations and improves overall throughput, as it obviates the need to run separate models for each task. As shown in prior work, this approach yields substantial resource savings compared with deploying separate networks per task [29].

Dataset

To train the calibration model, we use the SoccerNet 2023 dataset prepared for the SoccerNet 2023 challenge [30]. It contains over 21,000 frames captured by multiple synchronized broadcast cameras in professional soccer stadiums. Each frame includes semantic annotations of field markings (lines, penalty area, penalty spot, center circle) and rigid 3D landmarks (goalposts), represented as ordered 2D polylines derived from known 3D coordinates.

For reliable triangulation of annotated keypoints across viewpoints, a multi-camera configuration – typically 3–5 cameras simultaneously were employed. This allowed the camera parameters to generalize across different viewing angles and lens distortions. Moreover, the dataset's scale and diversity of stadiums, camera placements, and capture conditions exposed the model during training to a wide range of angles, zoom levels, and lighting, improving robustness to the variability of real-world broadcasts.

Specifically, this dataset was selected for the following characteristics:

- Point-based semantic labels of key field elements were produced independently of any specific camera parameters, enabling a unified representation that accommodates optical distortions and more complex geometric models.
- Synchronized viewpoints were used to impose geometric constraints across cameras – particularly re-projection consistency – providing additional accuracy control beyond single-view estimation.
- Annotations follow the ProCC protocol and cover both planar components of the field and non-planar landmarks (e.g., goalposts), enabling use in high-precision tasks such as out-of-play detection and 3D reconstruction.

Training on this multi-view dataset with detailed annotations yielded models capable of predicting camera parameters that accurately map the field's 3D geometry onto images captured by varied camera types and configurations [30].

Metrics

The Jaccard Index $\left(\mathrm{JaC}_\gamma\right)$ is used in this work to assess camera-calibration accuracy by comparing the projections of soccer field-marking segments with their annotations. Let s be the polyline obtained by projecting a field segment from the 3D model, and let $\hat{s}$ be the corresponding annotated polyline. Then the segment s is considered a true positive if, for all points $p \in s$,

$$\min_{q \in \hat{s}} d\,(p, q) < \gamma, \tag{1}$$

where $d(\cdot,\cdot)$ is the Euclidean distance in pixels, γ is the pixel error threshold, and q is a point on the annotated polyline $\hat{s}$, which serves as the reference projection. If no point satisfies this condition, the segment is counted as a false positive; segments present in the annotations but absent from the projections are counted as false negatives.

The Jaccard Index for threshold γ is defined as

$$\mathrm{JaC}_\gamma = \frac{\mathrm{TP}_\gamma}{\mathrm{TP}_\gamma + \mathrm{FP} + \mathrm{FN}}, \quad (2)$$

where TP_γ is the number of true-positive segments under threshold γ, FP is the number of false positives, and FN is the number of false negatives [30].

Solution completeness is evaluated via the completeness rate (CR), defined as the ratio of images for which camera parameters are successfully produced to the number of images containing at least four annotated field lines.

The final score (FS) is computed as the product of completeness and the Jaccard Index at $\gamma = 5$:

$$\mathrm{FS} = \mathrm{CR} \times \mathrm{JaC}_5. \quad (3)$$

This metric integrates accuracy and coverage of the camera (or homography) estimate, emphasizing that high accuracy (JaC) has limited value without sufficient coverage (CR).

Baseline Artificial Neural Network Architectures

Given the authors' constraints on computational resources, we selected the models W_{48}^{orig} and W_{48}^{dyn} as baselines, where the superscripts (orig) and (dyn) denote the original and dynamic versions of the model, respectively, and the number 48 indicates the model size.

The W_{48}^{orig} model was adopted from [24] without the final refinement module, since its use substantially slows processing. The architecture comprises two subnetworks: a keypoint-detection subnetwork trained for 200 epochs, and a line-recovery subnetwork trained for 100 epochs. Training time was assessed on an RTX 5090 GPU. On average, one training epoch for the keypoint subnetwork took about 25 minutes, while one epoch for the line subnetwork took about 100 minutes [24].

The W_{48}^{dyn} model was trained on the same data as W_{48}^{orig}, but with substantially fewer epochs: 29 epochs for the keypoint subnetwork and 22 epochs for the line subnetwork. This strategy enables an objective assessment of the proposed solution's performance dynamics relative to the most accurate approach, since fully training all experiments is highly resource intensive.

Thus, the chosen baseline models establish a starting point for subsequent comparisons in speed and accuracy. All performance measurements were conducted under identical hardware conditions: an NVIDIA RTX 4070 GPU and an Intel Core i9-14900K CPU.

The study results are presented in the tables 1 and 2.

Search for an Optimal Architecture

HRNet w48 Model

The HRNet w48 model, used in prior studies [24, 25], is the second-largest among the architectures considered and contains 77.5 million parameters. In the multi-view (MV) dataset setting without the PnL module, adopted from [24], the average frame-processing speed was measured at 8.56 FPS, which is insufficient for real-time applications. At the same time, the HRNet family includes lighter variants with fewer parameters. The FS metric for W_{48}^{orig} without the PnL module is 0.581, whereas for W_{48}^{dyn} it is 0.492.

The w48 model architecture is shown in Figure 1.

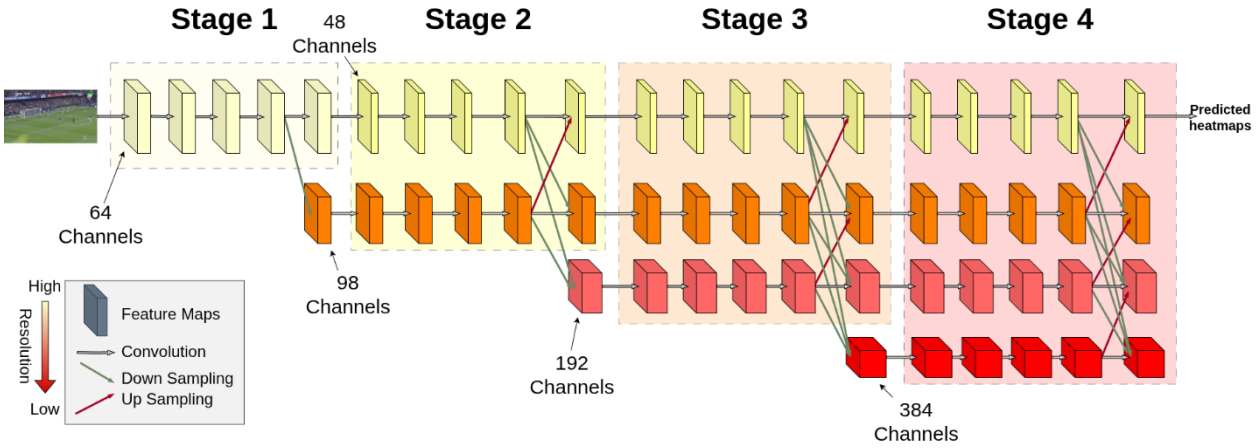

Fig.1. W_{48} model architecture

The W_{48} architecture consists of four sequential processing stages. In the first stage, the input image is transformed through a series of convolutional blocks into a 48-channel representation. In the subsequent stages (2–4), three parallel branches are employed with progressively reduced spatial resolution and increased channel counts: 98, 192, and 384 channels, respectively.

Feature exchange among these branches is performed via multi-branch down and up-sampling operations. Finally, features are aggregated, and the spatial resolution is gradually restored, producing output heatmaps for keypoints on the soccer field.

Application of Established Compact Architectures

In [31], the W_{48} models are presented alongside their simplified variants. From these, we selected the W_{32} and $W_{18\ \text{small}}$ architectures (hereafter W_{18}) for further study as the medium-size and smallest models, respectively. The simplification of these networks consists in reducing the number of blocks in the lower stages that process downsampled inputs, which substantially decreases the number of connections and model parameters.

The W_{32} architecture was trained for 22 epochs for the keypoint-detection task and 28 epochs for the line model, with a batch size of 4. Compared with W_{48}^{orig}, the FS metric decreased by 0.097, while relative to W_{48}^{dyn} – which was trained for approximately the same amount of time – the drop was only 0.008. The frame-processing speed increased by 4 FPS, reaching 12.5 FPS. This corresponds to a 150\% throughput gain, with less than a 1\% accuracy loss relative to W_{48}^{dyn} and about 10\% relative to W_{48}^{orig}.

The W_{18} architecture, which contains three times fewer parameters than W_{32} and six times fewer than W_{48}, delivered nearly double the frame-processing speed compared with W_{48}, reaching 18 FPS. This is 10 FPS higher than the original W_{48} (8.56 FPS) and 5.5 FPS higher than W_{32}. At the same time, accuracy decreased by 0.1 (10\%) compared with W_{48}^{orig} and by 0.01 (1\%) compared with W_{48}^{dyn}. Training for this architecture was conducted over 22 epochs for the keypoint model and 16 epochs for the line model.

The results obtained are presented in Tables 1 and 2 for the W_{32} and W_{18} models.

Design of the W_6 Architecture

To further optimize computational-resource usage, we propose the W_6 model, in which each processing stage contains a single block with increasing channel sizes: 6, 12, 24, and 48 at the final stage. The model architecture is shown in Figure 2. Here blocks from the original W_{48} that are omitted in W_6 are highlighted in gray.Training was conducted for 72 epochs for the keypoint-detection task and 17 epochs for the line model.

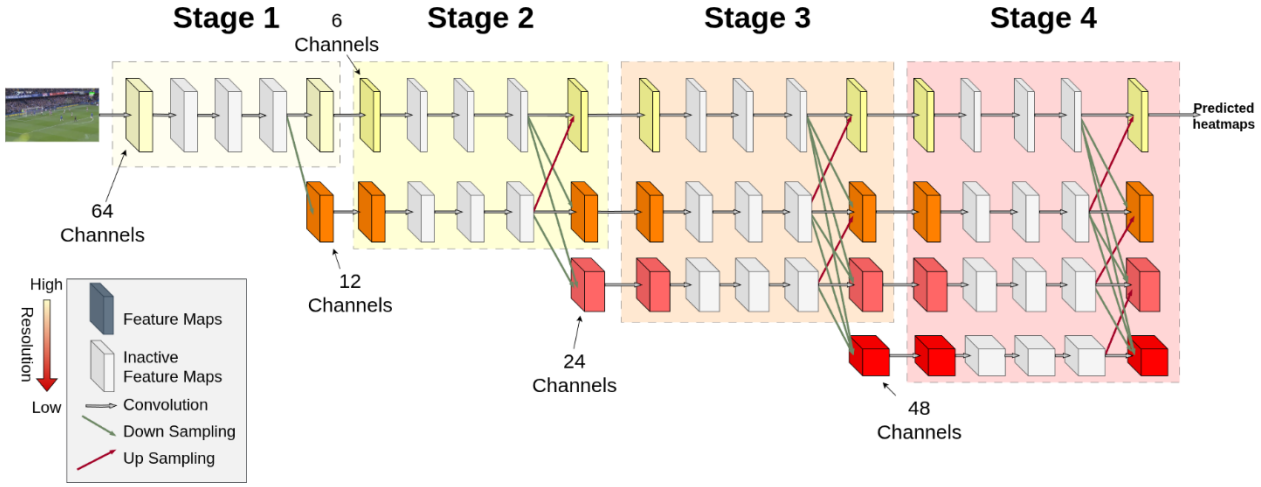

Fig. 2. Architecture of W_6 model

The W_6 architecture preserves the four-stage image-processing scheme with multi-branch downsampling and corresponding upsampling. However, instead of the channel sequence $48 \rightarrow 98 \rightarrow 192 \rightarrow 384$, it uses a compact sequence $6 \rightarrow 12 \rightarrow 24 \rightarrow 48$, which substantially reduces computational load.

Compared with the dynamic variant W_{48}^{dyn}, the proposed model achieved FS $= 0.455$ (a decrease of 0.037) with an average throughput of 22.973 FPS, which exceeds the baseline by 14.396 FPS. Relative to the original architecture W_{48}^{orig}, accuracy decreased by 0.126 (12.6%), while processing speed increased by 14.413 FPS to \$22.973\$ FPS (a 270% improvement).

These results indicate nearly a threefold increase in frame-processing speed compared with W_{48}^{orig}, with a moderate decrease in accuracy. The W_6 model appears promising for applications with strict real-time requirements in soccer analytics.

The obtained results are listed as W_6 in Tables 1 and 2.

Knowledge Distillation

To reduce the computational load of high-accuracy networks such as HRNet, we employ knowledge distillation [32]. In this paradigm, a more powerful teacher model imparts its expertise to a compact student model with minimal loss of accuracy. Within the HRNet framework, this approach preserves the multi-resolution, high-quality feature maps characteristic of the original architecture while simultaneously reducing parameter count and inference time [31].

We applied knowledge distillation to the W_{18} and W_6 architectures, selecting W_{48}^{orig} as the teacher due to its superior accuracy. The models were trained for an average of 35 epochs. For W_6, accuracy decreased from 0.455 to 0.422 FS. In contrast, W_{18} improved from 0.480 to 0.488, exceeding W_{48}^{dyn} by 0.006 FS and narrowing its gap to W_{48}^{orig} from 0.10 to 0.083.

The corresponding results are reported as W_{18}^{KD} and W_{6}^{KD} in Tables 1 and 2.

Combined Resulting Architecture

This work adopts a multi-task learning (MTL) approach, which enables training a single model to solve several related tasks simultaneously, fostering a shared data representation and improving the model's capacity to generalize.

In computer vision, this idea is typically implemented with an architecture that has a single backbone augmented with separate heads for each task. This design encourages the model to learn features that are jointly useful across tasks, reducing computational cost and increasing robustness to noise and limited data. For example, [29] successfully applied this approach to frames from a soccer match recording, where the model performed player re-identification, team-affiliation recognition, and role classification.

Here, we address two closely related subtasks—keypoint detection and line estimation. Because these tasks require similar internal representations, we employ a shared backbone with two heads: one for keypoints and one for lines.

For the MTL system, we evaluate two backbone variants — W_{18} and W_{48}. Results for different configurations are as follows:

- W_{48}: trained for 40 epochs. Throughput increased by 3 FPS with only a 0.08 FS (8%) decrease relative to W_{48}^{orig}, and a slight FS gain (+0.008, <1%) over W_{48}^{dyn}.
- W_{18} trained for 52 epochs. The final FS reached 0.484—less than 1% below W_{48}^{dyn} and 10% below W_{48}^{orig}. Throughput rose to 21 FPS (+12.5 FPS), making it comparable to the more compact W_6 model.

These outcomes achieve a favourable balance between accuracy and speed: depending on the chosen backbone, the model delivers either maximum accuracy (W_{48}) or high throughput and compactness (W_{18}).

The corresponding results are reported in Tables 1 and 2 as W_{18}^{MTL} and W_{48}^{MTL}.

Table 1

Comparative table of experimental metrics relative to the W_{48}^{dyn} model. KD — knowledge distillation; MTL — a model with a shared backbone and separate heads for each task.

Architecture	FS	FPS	Epochs (KP, Lines)
W_{48}^{dyn}	0.492	8.577	(29, 22)
W_{48}^{MTL}	0.500 (+0.008↑)	11.292 (+2.715↑)	40
W_{32}	0.484 (–0.008↓)	12.505 (+3.928↑)	(22, 28)
W_{18}	0.480 (–0.012↓)	17.746 (+9.169↑)	(22, 16)
W_{18}^{MTL}	0.484 (–0.008↓)	21.118 (+12.541↑)	52
W_{18}^{KD}	0.498 (+0.006↑)	18.474 (+9.897↑)	(37, 40)
W_{6}	0.455 (–0.037↓)	22.973 (+14.396↑)	(72, 17)
W_{6}^{KD}	0.422 (–0.070↓)	23.317 (+14.740↑)	(33, 23)

Table 2

Comparative table of experimental metrics relative to the W_{48}^{orig} model. KD — knowledge distillation; MTL — a model with a shared backbone and separate heads for each task.

Architecture	**FS**	**FPS**	**Epochs (KP, Lines)**
W_{48}^{orig}	0.581	8.560	(200, 100)
W_{48}^{MTL}	0.500 (–0.081↓)	11.292 (+2.732↑)	40
W_{32}	0.484 (–0.097↓)	12.505 (+3.945↑)	(22, 28)
W_{18}	0.480 (–0.101↓)	17.746 (+9.186↑)	(22, 16)
W_{18}^{MTL}	0.484 (–0.097↓)	21.118 (+12.558↑)	52
W_{18}^{KD}	0.498 (–0.083↓)	18.474 (+9.914↑)	(37, 40)
W_{6}	0.455 (–0.126↓)	22.973 (+14.413↑)	(72, 17)
W_{6}^{KD}	0.422 (–0.159↓)	23.317 (+14.757↑)	(33, 23)

Conclusions

This work focused on the runtime of the camera-calibration process and demonstrated that compact variants of HRNet can operate in real time with no substantial loss of accuracy.

Specifically, the W_6 architecture achieves roughly 23 FPS versus 8.6 FPS for the original W_{48}^{orig}, with FS scores of 0.455 and 0.581, respectively. The W_{18}^{MTL} variant reaches 21.1 FPS with an FS of 0.484. Using knowledge distillation, the W_{18}^{KD} variant improves to FS 0.498 at 18.5 FPS, whereas the ultra-compact W_{6}^{KD} shows a decrease to FS 0.422.

Relative to the partially fine-tuned dynamic model W_{48}^{dyn} (a fairer comparator, since not all variants were fully retrained), the lighter W_{18}^{MTL} and W_6 architectures deliver comparable or slightly lower FS values while offering substantially higher throughput, whereas W_{48}^{orig} remains the upper bound on accuracy.

Overall, the proposed family of neural-network architectures enables selection of an optimal trade-off between accuracy and speed for camera calibration, subject to specific hardware constraints in soccer analytics.

References

1. Blanchard N., Skinner K., Kemp A., Scheirer W., Flynn P. «Keep me in, coach!»: A computer vision perspective on assessing ACL injury risk in female athletes. 2019. https://doi.org/10.1109/WACV.2019.00150.

2. Wang Z., Veličković P., Hennes D., Tomašev N., Prince L., Kaisers M., Bachrach Y., Elie R., Wenliang L.K., Piccinini F. TacticAI: an AI assistant for football tactics // Nature Communications. 2024. Vol. 15, No. 1. https://doi.org/10.1038/s41467-024-45965-x.

3. Perez-Yus A., Agudo A. Matching and recovering 3D people from multiple views // Proc. 2022 IEEE/CVF Winter Conf. Applications of Computer Vision (WACV). 2022. https://doi.org/10.1109/WACV51458.2022.00125.

4. Cioppa A., Deliège A., Magera F., Giancola S., Barnich O., Ghanem B., Van Droogenbroeck M. Camera calibration and player localization in SoccerNet-v2 and investigation of their representations for action spotting // Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW). 2021. https://doi.org/10.1109/CVPRW53098.2021.00511.

5. Manafifard M. A review on camera calibration in soccer videos // Multimedia Tools and Applications. 2023. https://doi.org/10.1007/s11042-023-16145-8.

6. Cuevas C., Quilón D., García N. Automatic soccer field of play registration // Pattern Recognition. 2020. https://doi.org/10.1016/j.patcog.2020.107278.

7. Giancola S., Cioppa A., Deliège A., Magera F., Somers V., Kang L., Zhou X., Barnich O., De Vleeschouwer C., Alahi A. SoccerNet 2022 challenges results. 2022.

8. Theiner J., Ewerth R. TVCalib: Camera calibration for sports field registration in soccer. 2023.

9. Strange W. Strange technology allows us all to train like champions // Planet Innovation News. 2015.

10. Fleet E. How much does sport data cost? // Stats Perform Blog. 2023.

11. Sasikala P., Pragathi V., Dharshini N.S.P., Sreeranjani P., Swathi A.K. Sports analysis software for football education and intellectual property awareness // International Journal of Innovative Research in Technology. 2025. Vol. 11, No. 12.

12. Torres-Ronda L. et al. Tracking systems in team sports: a narrative review of applications of the data and sport specific analysis // Sports Medicine – Open. 2022. Vol. 8, No. 15. https://doi.org/10.1186/s40798-022-00408-z.

13. Csanalosi G., Dobreff G., Pašić A., Molnár M., Toka L. Low-cost optical tracking of soccer players // Machine Learning and Data Mining for Sports Analytics (ECML PKDD 2020 Workshop). 2020.

14. Mavrogiannis P., Maglogiannis I. Amateur football analytics using computer vision // Neural Computing and Applications. 2022. Vol. 34, No. 22, pp. 19639–19654. https://doi.org/10.1007/s00521-022-07692-6.

15. Lowe D.G. Distinctive image features from scale-invariant keypoints // International Journal of Computer Vision. 2004. https://doi.org/10.1023/B:VISI.0000029664.99615.94.

16. Goorts P., Maesen S., Liu Y., Dumont M., Bekaert P., Lafruit G. Self-calibration of large scale camera networks. // IEEE. 2014.

17. Niu Z., Gao X., Tian Q. Tactic analysis based on real-world ball trajectory in soccer video // Pattern Recognition. 2012. https://doi.org/10.1016/j.patcog.2011.10.023

18. Bu J., Lao S., Bai L. Automatic line mark recognition and its application in camera calibration in soccer video. // IEEE. 2011.

19. Homayounfar N., Fidler S., Urtasun R. Sports field localization via deep structured models // arXiv preprint arXiv:1707.03876. 2017.

20. Citraro L., Márquez-Neila P., Savarè S., Jayaram V., Dubout C., Renaut F., Hasfura A., Ben Shitrit H., Fua P. Real-time camera pose estimation for sports fields // Machine Vision and Applications. 2020. Vol. 31, No. 3. https://doi.org/10.1007/s00138-020-01064-7.

21. Jiang W., Higuera J.C.G., Angles B., Sun W., Javan M., Yi K.M. Optimizing through learned errors for accurate sports field registration. 2020.

22. Tarashima S. SFLNet: direct sports field localization via CNN-based regression // ACPR. 2020.

23. Theiner J., Ewerth R. TVCalib: Camera calibration for sports field registration in soccer // Proc. 2008 IEEE Winter Conf. Applications of Computer Vision (WACV). 2008. https://doi.org/10.48550/arXiv.2207.11709.

24. Gutiérrez-Pérez M., Agudo A. PnLCalib: Sports Field Registration via Points and Lines Optimization // arXiv preprint arXiv:2404.08401. 2024.

25. Falaleev N.S., Chen R. Enhancing soccer camera calibration through keypoint exploitation // Proc. 7th ACM Int. Workshop on Multimedia Content Analysis in Sports. 2024.

26. Gutiérrez-Pérez M., Agudo A. No bells just whistles: Sports field registration by leveraging geometric properties // Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR). 2024. pp. 3325–3334. https://doi.org/10.48550/arXiv.1908.07919.

27. Yu C., Xiao B., Gao C., Yuan L., Zhang L., Sang N., Wang J. Lite-HRNet: A lightweight high-resolution network // Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR). 2021. pp. 10440–10450. https://doi.org/10.1109/CVPR46437.2021.01012.

28. Bose S., Sarkar S., Chakrabarti A. SoccerKDNet: a knowledge distillation framework for action recognition in soccer videos // Int. Conf. Pattern Recognition and Machine Intelligence. 2023.

29. Mansourian A.M., Somers V., De Vleeschouwer C., Kasaei S. Multi-task learning for joint re-identification, team affiliation, and role classification for sports visual tracking // Proc. 6th Int. Workshop on Multimedia Content Analysis in Sports. 2023.

30. Magera F., Hoyoux T., Barnich O., Van Droogenbroeck M. A universal protocol to benchmark camera calibration for sports // Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR). 2024. pp. 3335–3346.

31. Wang J., Sun K., Cheng T., Jiang B., Deng C., Zhao Y., Liu D., Mu Y., Tan M., Wang X. et al. Deep high-resolution representation learning for visual recognition // IEEE Trans. Pattern Anal. Mach. Intell. 2020. Vol. 43, No. 10, pp. 3349–3364. https://doi.org/10.1109/TPAMI.2020.2983686.

32. Hinton G., Vinyals O., Dean J. Distilling the knowledge in a neural network // arXiv preprint arXiv:1503.02531. 2015.