Toward football analytics via AI

Joonghyun Bae

April 2024

1 Toward football analytics via AI

People love football and I am one of football fans. For football fans, it is tempting to have visual football analytics of their favorite football clubs as in the Match Of The Day(famous BBC TV Show for match analysis). However, the only data available for fans are football broadcast videos, and the well-polished match data and its visualizations are often expensive and inaccessible to public. Motivated by recent researches on football analytics using AI [HFU17, NCH21, CSH$^+$22, OHG$^+$22, MODP23a, WVH$^+$23], I have been working on constructing AI model for football data visualization. More precisely, the goal of this project is to make a "minimap" version of short football broadcast videos. This article presents the partial result toward this goal.

2 Football field registration

A (two-dimensional) homography matrix is an invertible matrix $H : \mathbb{R}^3 \to \mathbb{R}^3$ whose last entry is equal to 1:

$$H = \begin{pmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & 1 \end{pmatrix},$$

where $h_i \in \mathbb{R}$ for $i = 1, \dots, 8$. Let Image_j denote an image for $j = 1, 2$ and let $(x, y), (X, Y)$ denote its pixel-coordinates, respectively. A homography matrix H defines a bijective transformation

$$\varphi_H : \text{Image}_1 \to \text{Image}_2, \quad (x, y) \mapsto (X, Y) := \varphi_H(x, y)$$

by the condition

$$(cX, cY, c) = H(x, y, 1), \quad c = h_7x + h_8y + 1.$$

Any pair of two images of the same planar surface in space is related by this transformation. In other words, it describes a change of camera view of images for planar objects, see [HZ03] for details.

Given a broadcast image of a football match, the goal of football field registration is to find a homography matrix which describes a change of camera view of the broadcast image to bird's-eye view for football field. This is the first step towards the precise analysis of football matches based on broadcast videos since any numerical analysis of football matches should be based on a fixed coordinate system of football fields.

2.1 Football field registration as image segmentation

Let $(x_i, y_i) \in \text{Image}_1$ and $(X_i, Y_i) \in \text{Image}_2$, for $i = 1, \dots, 4$, be pairwise distinct points in two images, respectively. For a generic choice of correspondences $(x_i, y_i) \leftrightarrow (X_i, Y_i)$, the Direct Linear Transformation algorithm or the DLT algorithm finds a unique homography matrix H whose induced transformation $\varphi_H : \text{Image}_1 \to \text{Image}_2$ satisfies

$$\varphi_H(x_i, y_i) = (X_i, Y_i), \quad \text{for } i = 1, \dots, 4.$$

In degenerate cases, the DLT algorithm finds a homography matrix which minimizes the sum of L^2-distances between $\varphi_H(x_i, y_i)$ and (X_i, Y_i) for $i = 1, \dots, 4$. (See [HZ03] for details of the DLT algorithm.) By means of the DLT algorithm, a problem of finding a homography matrix reduces to a problem of finding (at least) four pairs of corresponding points in two images.

Figure 1: A football broadcast image and its bird's-eye view transformed image

Recent researches on football field registration in this direction find pairs of corresponding points in football broadcast image and the football field template image via deep neural networks, see [NCH21, CSH+22, MODP23b]. More precisely, those works define an equi-distributed collection of keypoints in the football template image or the "minimap". Then, they train neural networks to predict the pixel coordinates of the "imaginary" keypoints in the football broadcast image, which are images of keypoints in the minimap under the (inverse) transformation of the estimated homography matrix. See Figure 2 below for equi-distributed keypoints in the minimap and the visualization of "imaginary" keypoints in the football braodcase image. To achieve robustness of the homography matrix estimation, the RANdom SAmple Consensus or the RANSAC algorithm [FB81] is applied.

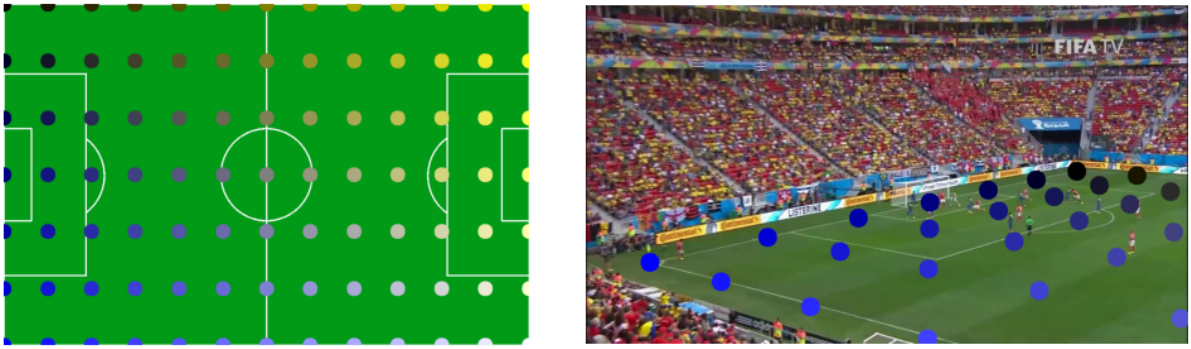

Figure 2: Keypoints in the minimap and the groundtruth keypoints projected via homography matrix

2.2 Result and discussion

Motivated by recent researches, we have trained a deep neural network model for image segmentation called the Dense Prediction Transformers or the DPT [RBK21] to predict the projected keypoints in the broadcast football images. We finetuned the DPT model for semantic segmentation pretrained on ADE20k dataset [ZZP+17] adapting the finetuning method suggested by LoRA [HSW+21]. Then, we estimated a homography matrix by the DLT algorithm together with the RANSAC, whose induced transformation provides the bird's-eye view of the football broadcast images. Finally, we made the minimap image representing the region viewed in the football broadcast image. See Figure 3 below for a few sample images showing our results.

To evaluate the accuracy of the estimated homography $\hat{H}$, we used an error function E defined by

$$\mathrm{E}(\hat{H}) := \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} d_{\mathrm{std}}(x, \varphi_{\hat{H}} \circ \varphi_{H^{-1}}(x)),$$

where $\mathcal{S}$ is the set of keypoints x in the minimap such that $\varphi_{H^{-1}}(x)$ is in the broadcast image for the groundtruth homography matrix H and d_{std} denotes the standard Euclidean distance. The error function E computes a rough

Figure 3: Left column: test images. Middle column: predicted keypoints. Right column: estimated minimaps

estimation of the average error, with respect to the distance, of each point in the minimap viewed in the broadcast image. Throughout the validation dataset, the average value of $\mathrm{E}(\hat{H})$ is 0.775 (yards). In other words, the estimated coordinates of the minimap is accurate up to 0.7 meter in average.

The DLT algorithm fails to predict a homography matrix if the number of predicted keypoints is less than four. Throughout the validation dataset, the average number of predicted keypoints is 18.8. and the minimum number of predicted keypoints is 7, which is strictly larger than 4. This result supports the robustness of the homography estimation based on keypoints prediction.

The average inference time for the homography estimation on the validation dataset is 275ms per frame with an Intel i9-10940X CPU and a single Nvidia GeForce RTX 3090 GPU. Since the ultimate goal of this project needs to estimate the homography matrix for each frame in football broadcast videos, faster inference time without losing accuracy is required. This aspect will be studied in the future.

2.3 Implementation details

Data

Among a few public-available dataset([HFU17], [CSH$^+$22]), we used the TS-Worldcup dataset [CSH$^+$22] consisting of a number of 2925 football broadcast images and the corresponding groundtruth homography matrices for train dataset and a number of 887 images and homography matrices for test dataset. The resolution of football broadcast images is 1280×720. The transformations induced by groundtruth homography matrices describes the change of camera view from football broadcast images to the bird's-eye view. The number of keypoints in the minimap is 91 consisting of equally distributed keypoints with 7 rows and 13 columns. We expanded each projected keypoint in the broadcast image to a disk of pixel-radius 20 to make groundtruth segmentation masks.

Model

For a semantic segmentation model to detect the projected keypoints, we used the deep neural network called the Dense Prediction Transformers or the DPT. The principal building block of the DPT is the transformer module,

whose construction is motivated by the transformer models introduced by [VSP+17] and applied to vision task by [DBK+20]. Among many other properties, (vision) transformer processes representations of an input image at a constant and relatively high resolution and has a global receptive field at every stage. These properties allow the DPT to provide finer-grained and more globally coherent predictions, which are suitable for the keypoints detection. We used the DPT model pretrained with the ADE20K dataset and tailored the last two layers to predict 92 classes(91 keypoints together with the background class). The number of parameters of the used model is 343M. The input image resolution is 480×480 and the output is the predicted segmentation mask of resolution 480×480. For the loss function, we used the cross-entropy loss with weight 1 for background classes and 10 for keypoints classes. The average inference time for the model inference on the validation dataset is 27ms per frame on a single Nvidia GeForce RTX 3090 GPU.

Data augmentation and preprocessing

We applied the following data augmentation to the train dataset. We normalized the RGB value with mean $0.485, 0.456, 0.406$ and standard deviation $0.229, 0.224, 0.225$, respectively. We randomly erased at most 20 small rectangles in the image and replaced its pixel-values to the constant $(127.5, 127.5, 127.5)$. We randomly flipped the images horizontally with probability 0.5. We note that the corresponding segmentation masks also have to be transformed as shown in Figure 4 below in contrast to the usual semantic segmentation tasks. We randomly cropped the images with probability 0.5 and with the random scale in range 0.7 to 1. Then, we resized the input image to the size 480×480 with bicubic interpolation.

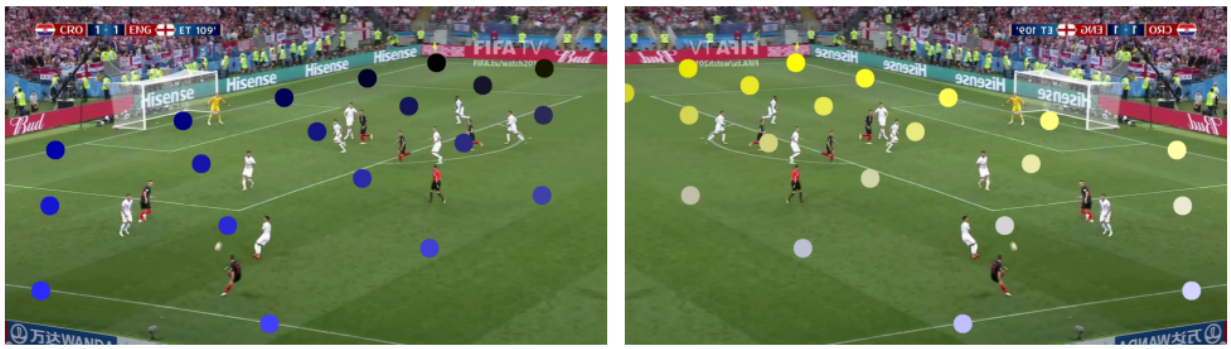

Figure 4: Left: a visualization of an input image together with its segmentation label, Right: the horizontally flipped image together with the segmentation label adjusted to the horizontal flip.

Training

We finetuned the pretrained DPT model using the Low-Rank Adaptation or the LoRA [HSW+21]. Instead of fine-tuning all the parameters of large models, LoRA only adapts a small set of additional parameters that interact with the pretrained weights through low-rank matrices. It significantly reduces the number of parameters to be updated and offers faster convergence as well as computational efficiency. We applied LoRA to attention layers and linear layers to reduce the trainable parameters to 2.5% of the original model: the number of original parameters is 343M and the number of trained parameters using LoRA is 8.5M. The rank of LoRA matrix was set to 16 and the alpha parameter is set to 16. We finetuned the model for 32 epochs with 4 images per batch. We used the AdamW optimizer [LH17] with the default parameters in PyTorch. We used the step learning rate scheduler with the starting learning rate 10^{-4}, step size 10, and the decaying parameter 0.5.

Homography estimation

To estimate the homography matrix, we computed the center of mass of each keypoints on segmentation mask given by the output of model. Then, we applied the DLT algorithm together with RANSAC to obtain the estimated homography matrix. The RANSAN reprojection error threshold was set to 8.

3 Ongoing project: Player tracking

3.1 First step: Player detection

As the first step toward player tracking for football broadcast videos, it is necessary to detect players and a ball in an image. We used the deep neural network designed to detect objects in images, called the YOLOS [FLW+21]. We trained the YOLOS model to predict players, goalkeepers, referees, and balls in football broadcast images. Combining with football field registration technique, we visualized the positions of players in the minimap images.

3.2 Result

The model predicts bounding boxes and the confidence scores of the detected objects in the images. We detect players, goalkeepers, referees, balls, and unknown figures(for example, coaches or team doctors). A few sample detection results for test dataset can be seen in Figure 5 below. Combined with the football field registration, the football broadcast images can be transformed to the corresponding minimaps with the location of detected players depicted on the minimaps, see Figure 6 below.

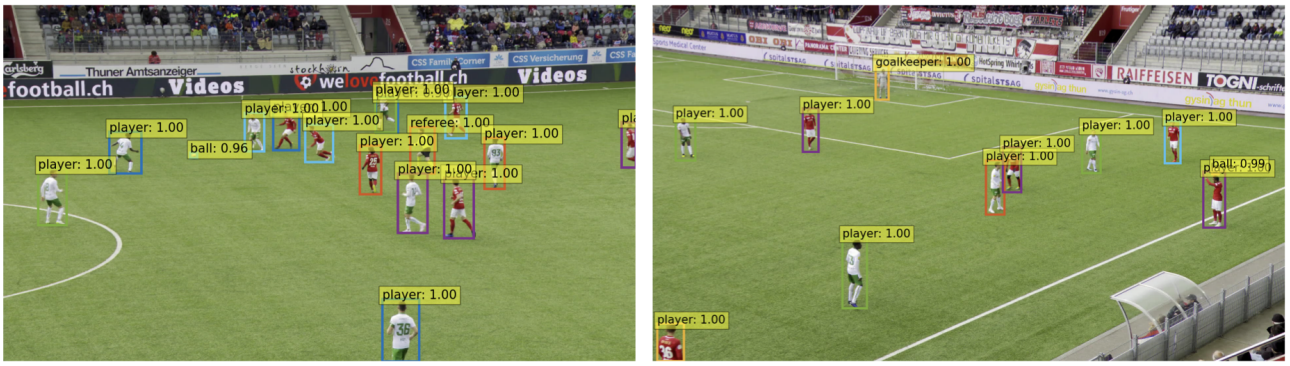

Figure 5: Sample detection results : visualization of the bounding boxes of detected objects together with classes and confidence scores.

3.3 Implementation details

Data

We used the public-available SoccerNet dataset for tracking [DCG+20]. This dataset consists of a number of 42750 football broadcast images with COCO-style tracking labels for train dataset and a number of 36750 images with tracking labels for test dataset.

Model

Object detection model we used is YOLOS[FLW+21] whose network design is a reminiscent of the DEtection TRansformers or the DETR [CMS+20]. It consists of transformer-based backbone called the DeiT [TCD+20] together with transformer encoder to predict the coordinates of bounding boxes and the confidence scores of the detected objects in images. As in the case of [CMS+20], its loss function is the weighted sum of cross-entropy loss and the generalized bounding box loss, where the latter is the weighted sum of the GIOU(generalized intersection over union) loss and the L_1 loss for bounding boxes. We used the base-sized DeiT pretrained on the Imagenet-1k [RDS+15] as a backbone. The number of parameters of the model is 127M.

Training

We trained the model for 10 epochs with 1 image per batch. We used the AdamW optimizer with weight decay parameter 10^{-4} and the cosine annealing learning rate scheduler with initial learning rate 2.510^{-5} and the maximum iteration parameter 10. The maximum norm for gradient clipping was 10.

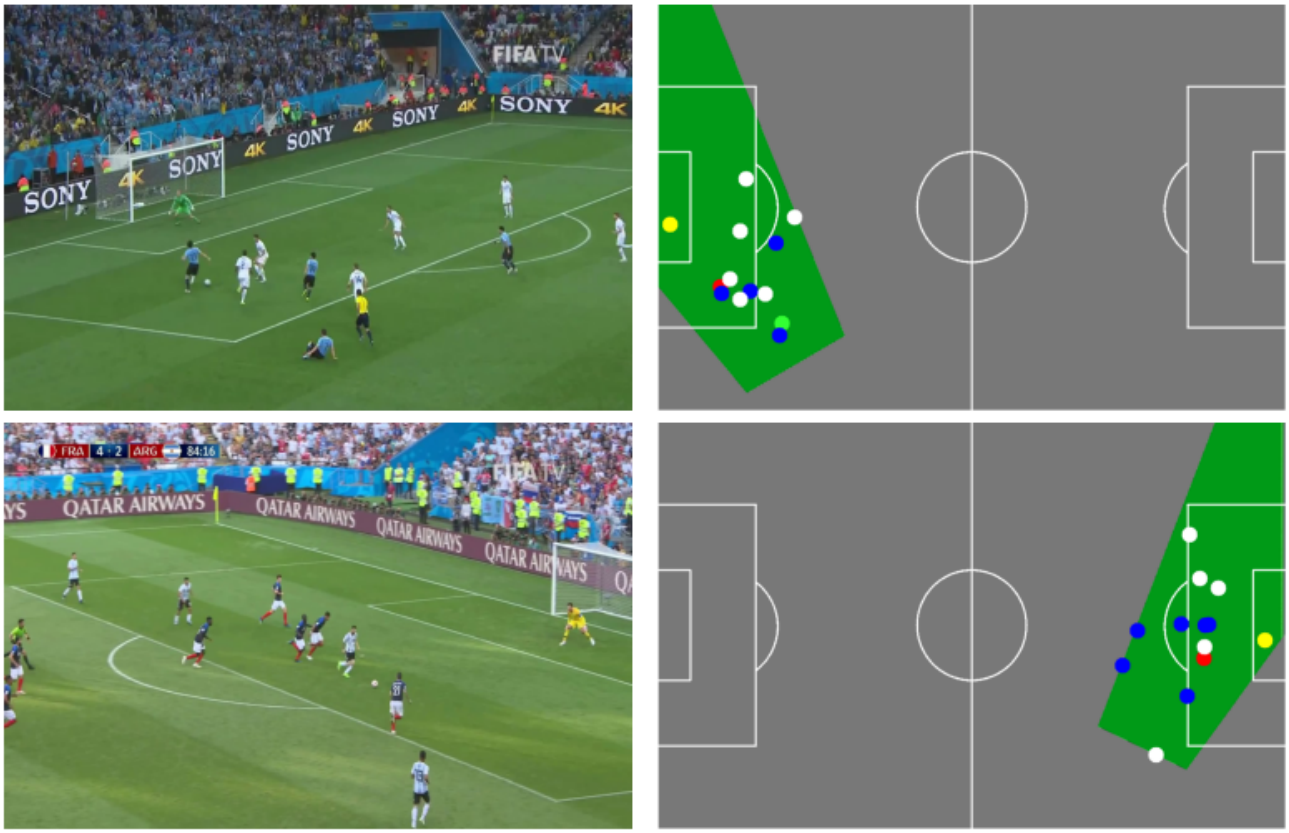

Figure 6: Groundtruth images(left) and transformed minimaps(right). White and blue dots represent players, yellow dot represent goalkeepers, red dot represent balls, and green dots represent referees.

3.4 Second step: Tracking detected players

We plan to apply the OC-SORT algorithm [CPW+23] for tracking detected players. The OC-SORT algorithm is a detection-based tracking algorithm whose advantages are robustness during occlusion and non-linear motion.

References

[CMS+20] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, *End-to-end object detection with transformers*, CoRR **abs/2005.12872** (2020).

[CPW+23] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani, *Observation-centric sort: Rethinking sort for robust multi-object tracking*, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 9686–9696.

[CSH+22] Yen-Jui Chu, Jheng-Wei Su, Kai-Wen Hsiao, Chi-Yu Lien, Shu-Ho Fan, Min-Chun Hu, Ruen-Rone Lee, Chih-Yuan Yao, and Hung-Kuo Chu, *Sports field registration via keypoints-aware label condition*, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2022, pp. 3522–3529.

[DBK+20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, *An image is worth 16x16 words: Transformers for image recognition at scale*, CoRR **abs/2010.11929** (2020).

[DCG+20] Adrien Deliège, Anthony Cioppa, Silvio Giancola, Meisam Jamshidi Seikavandi, Jacob V. Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B. Moeslund, and Marc Van Droogenbroeck, *Soccernet-v2 : A dataset and benchmarks for holistic understanding of broadcast soccer videos*, CoRR **abs/2011.13367** (2020).

[FB81] Martin A Fischler and Robert C Bolles, *Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography*, Communications of the ACM **24** (1981), no. 6, 381–395.

[FLW+21] Yuxin Fang, Bencheng Liao, Xinggang Wang, Jiemin Fang, Jiyang Qi, Rui Wu, Jianwei Niu, and Wenyu Liu, *You only look at one sequence: Rethinking transformer in vision through object detection*, CoRR **abs/2106.00666** (2021).

[HFU17] Namdar Homayounfar, Sanja Fidler, and Raquel Urtasun, *Sports field localization via deep structured models*, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4012–4020.

[HSW+21] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen, *Lora: Low-rank adaptation of large language models*, CoRR **abs/2106.09685** (2021).

[HZ03] Richard Hartley and Andrew Zisserman, *Multiple view geometry in computer vision*, Cambridge university press, 2003.

[LH17] Ilya Loshchilov and Frank Hutter, *Decoupled weight decay regularization*, arXiv preprint arXiv:1711.05101 (2017).

[MODP23a] Adrien Maglo, Astrid Orcesi, Julien Denize, and Quoc Cuong Pham, *Individual locating of soccer players from a single moving view*, Sensors **23** (2023), no. 18.

[MODP23b] ———, *Individual locating of soccer players from a single moving view*, Sensors **23** (2023), no. 18, 7938.

[NCH21] Xiaohan Nie, Shixing Chen, and Raffay Hamid, *A robust and efficient framework for sports-field registration*, 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1935–1943.

[OHG+22] Shayegan Omidshafiei, Daniel Hennes, Marta Garnelo, Zhe Wang, Adria Recasens, Eugene Tarassov, Yi Yang, Romuald Elie, Jerome T Connor, Paul Muller, et al., *Multiagent off-screen behavior prediction in football*, Scientific reports **12** (2022), no. 1, 8638.

[RBK21] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun, *Vision transformers for dense prediction*, CoRR **abs/2103.13413** (2021).

[RDS+15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., *Imagenet large scale visual recognition challenge*, International journal of computer vision **115** (2015), 211–252.

[TCD+20] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou, *Training data-efficient image transformers & distillation through attention*, CoRR **abs/2012.12877** (2020).

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, *Attention is all you need*, CoRR **abs/1706.03762** (2017).

[WVH+23] Zhe Wang, Petar Veličković, Daniel Hennes, Nenad Tomašev, Laurel Prince, Michael Kaisers, Yoram Bachrach, Romuald Elie, Li Kevin Wenliang, Federico Piccinini, William Spearman, Ian Graham, Jerome Connor, Yi Yang, Adrià Recasens, Mina Khan, Nathalie Beauguerlange, Pablo Sprechmann, Pol Moreno, Nicolas Heess, Michael Bowling, Demis Hassabis, and Karl Tuyls, *Tacticai: an ai assistant for football tactics*.

[ZZP+17] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba, *Scene parsing through ade20k dataset*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 633–641.