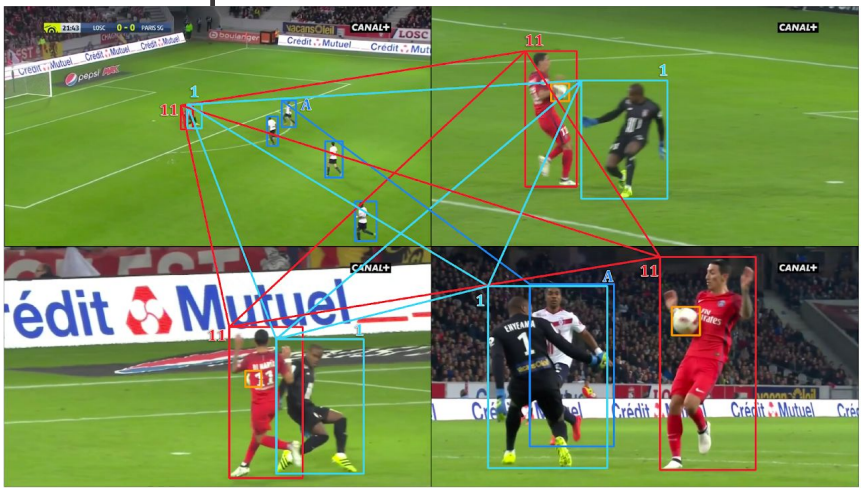
Domain-guided Masked Autoencoders for Unique Player Identification

Bavesh Balaji, Jerrin Bright, Sirisha Rambhatla, Yuhao Chen,
Alex Wong, John Zelek and David Clausi
Vision and Image Processing Lab,
University of Waterloo,
Waterloo, Ontario, Canada

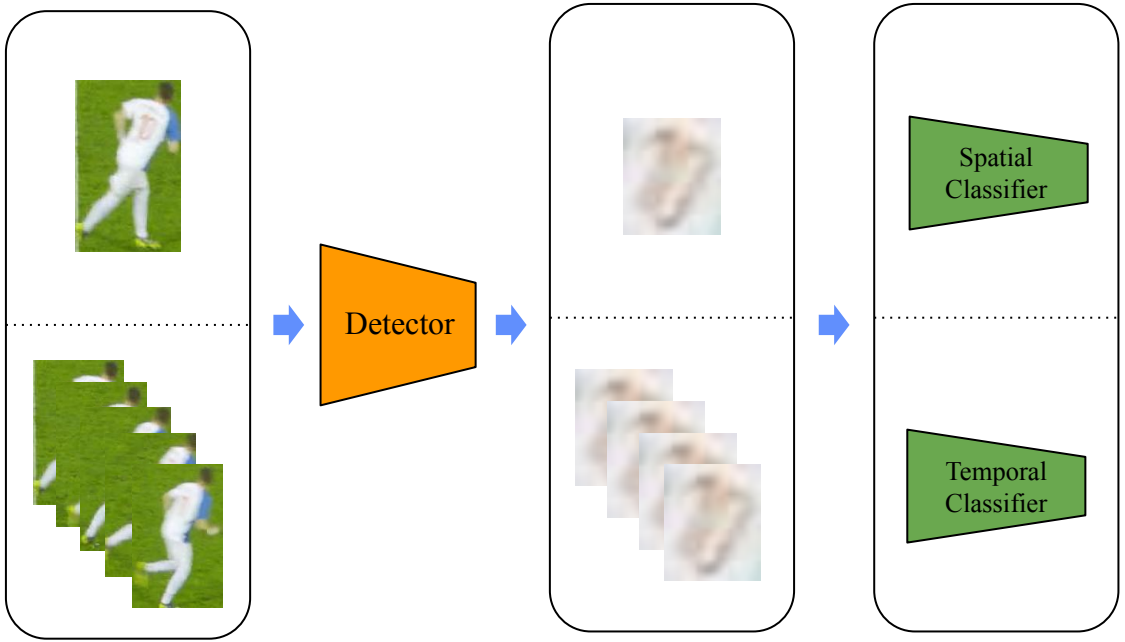
MOTIVATION

- Jersey number recognition – Common approach for player identification.
- In-game analytics, enhanced broadcast experience.



EXISTING WORKS

- Formulate as a classification problem.
  - Most methods operate on static images[1, 2].
    - Do not consider temporal aspect.
    - Datasets created in isolated environments.
  - Few works use tracklets[3, 4].
    - Consider temporal aspect.



[1] D. Bhargavi, E. P. Coyotl, and S. Gholami, “Knock, knock. who’s there? – identifying football player jersey numbers with synthetic data,” 2022.
[2] G. Li, S. Xu, X. Liu, L. Li, and C. Wang, “Jersey number recognition with semi-supervised spatial transformer network,” in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1864–18647, 2018.
[3] K. Vats, W. J. McNally, P. Walters, D. A. Clausi, and J. S. Zelek, “Ice hockey player identification via transformers and weakly supervised learning,” 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).
[4] A. Chan, M. D. Levine, and M. Javan, “Player identification in hockey broadcast videos,” Expert Syst. Appl., vol. 165, p. 113891, 2020.

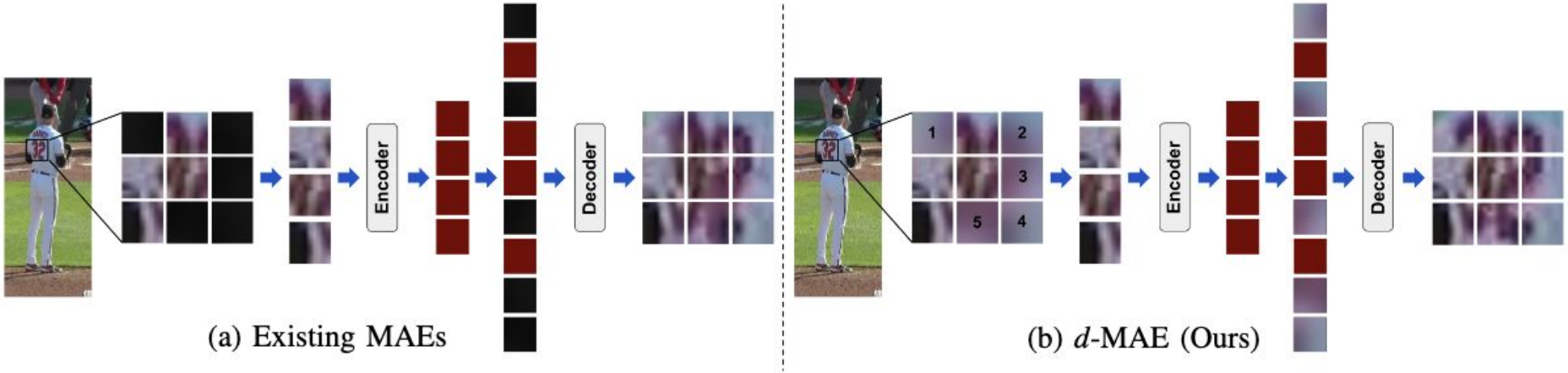
LIMITATIONS

- Prone to **motion blur** & occlusions.
- Existing Spatial feature extractors are not robust.
- JN not visible in most frames.

Motion Blur

Occlusion

d-MAE

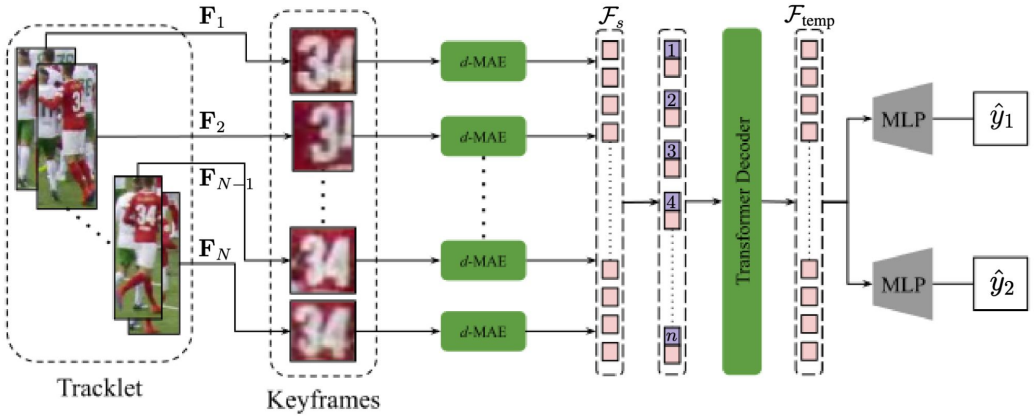


(a) Existing MAEs

(b) *d*-MAE (Ours)

- Existing MAEs completely black-out random subset of image patches.
- We introduce motion blur artifacts on random patches.

OUR METHOD



The proposed approach comprises several key steps:

1. Keyframe Identification: Each input tracklet is passed through the KfID module which identifies keyframes that contain high-level context of the jersey number, and localizes it.

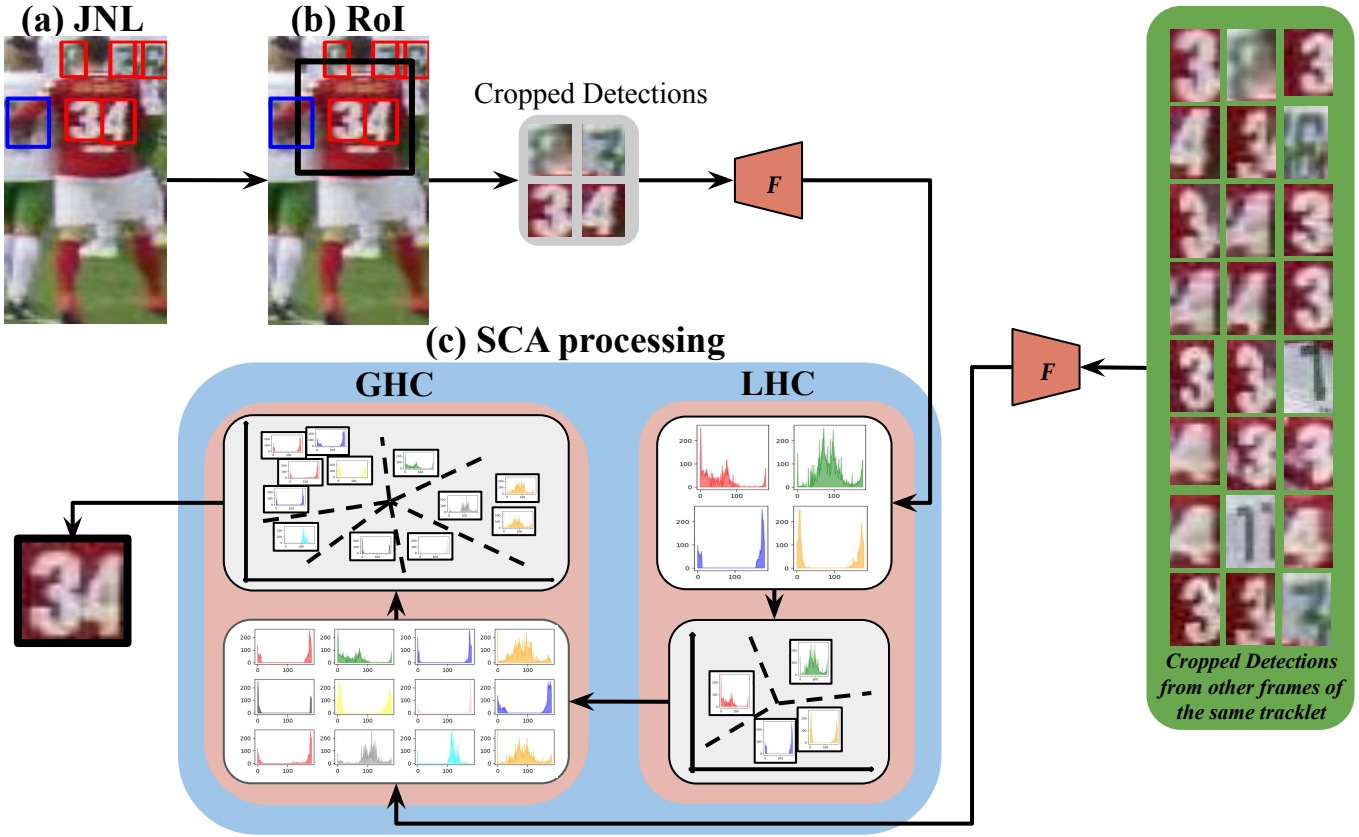
2. *d*-MAE: The extracted frames are then individually passed through our proposed *d*-MAE to extract the spatial features $\mathcal{F}_s$ of each keyframe.

3. Temporal Transformer Decoder: The extracted spatial features $\mathcal{F}_s$ are passed through a transformer network to extract the temporal features $\mathcal{F}_{\text{temp}}$ necessary to identify the jersey number reliably.

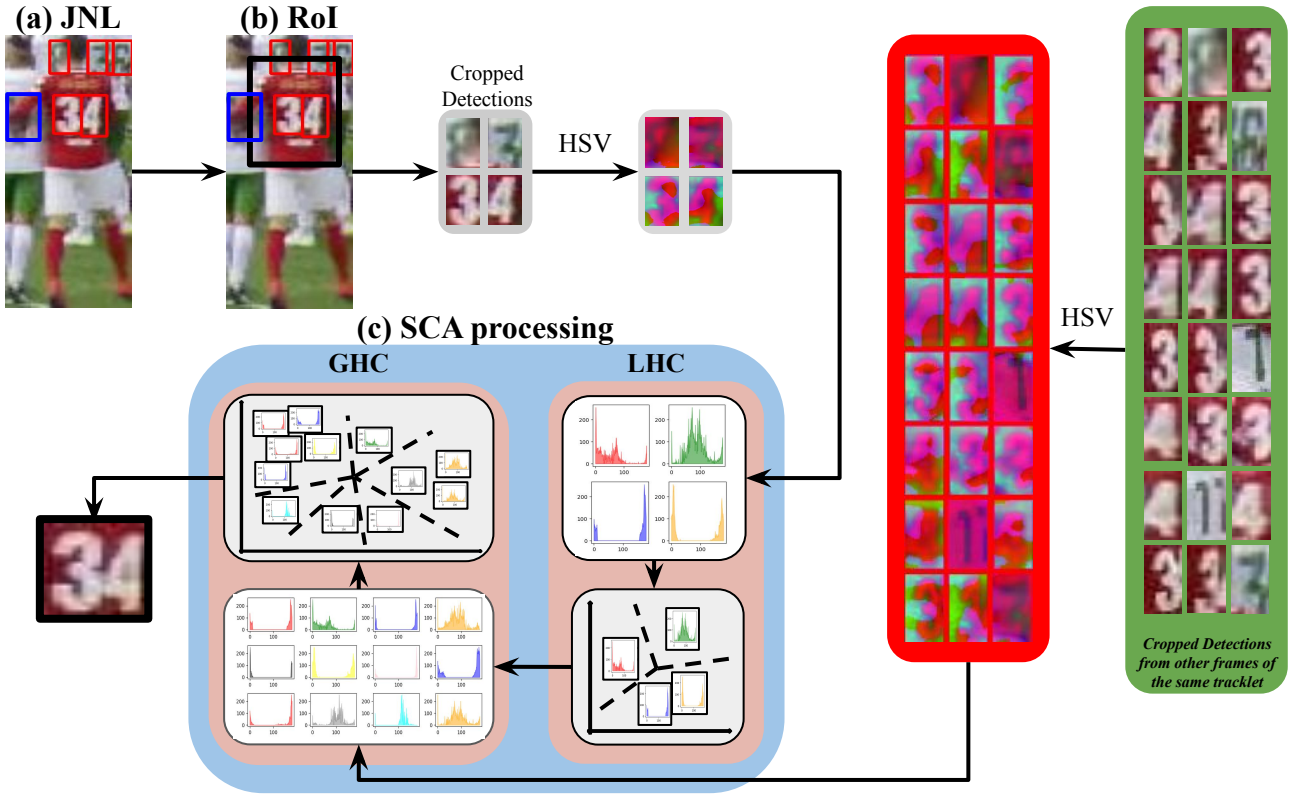
LOSS FUNCTIONS

- MAE loss
  - Reconstruction loss: $||\hat{\mathbf{I}} - \mathbf{I}||_2$
  - Siamese Loss:- $\mathcal{L}_{\text{siamese}} = ||h(\hat{\mathbf{I}}) - h(\mathbf{I})||_1$
  - $\mathcal{L}_{\text{mae}} = \sigma_1 \ ||\hat{\mathbf{I}} - \mathbf{I}||_2 + \sigma_2 \ \mathcal{L}_{\text{siamese}}$
- Multi-task classifier loss
  - $\mathcal{L}_{\text{class}} = -\sum_{i=0}^{10} y_1^i \log \hat{y}_1^i - \sum_{j=0}^{10} y_2^j \log \hat{y}_2^j$

KfID MODULE


(a) JNL
(b) RoI
Cropped Detections
F
(c) SCA processing
GHC
LHC
F
Cropped Detections from other frames of the same tracklet

KfID MODULE


(a) JNL
(b) RoI
Cropped Detections
HSV
(c) SCA processing
GHC
LHC
HSV
Cropped Detections from other frames of the same tracklet

DATASETS

Table I: Dataset split-up for training, validation and testing.

	SoccerNet Dataset			Ice Hockey Dataset			Baseball Dataset		
Dataset	**Tracklets**	**Images**	**Keyframes**	**Tracklets**	**Images**	**Keyframes**	**Tracklets**	**Images**	**Keyframes**
Train	1,141	587,543	71,021	2,829	540,339	162,101	105	21,050	18,344
Validation	286	146,886	19,609	176	33,616	11,084	15	2,962	2,571
Test	1,211	565,758	70,445	505	96,455	28,937	30	5,988	4,640
Challenge	1,426	750,092	101,307	-	-	-	-	-	-
Total	4,064	2,052,306	262,382	3,510	670,410	202,122	150	30,000	25,555

RESULTS

Table II: Quantitative comparison of our model with the state-of-the-art on the three datasets.

	SoccerNet		Ice Hockey	Baseball
Method	**Test Acc**	**Challenge Acc**	**Test Acc**	**Test Acc**
Gerke et al [28]	32.57	35.79	61.20	64.47
Vats et al [8]	46.73	49.88	83.17	87.61
Li et al [4]	47.85	50.60	81.15	88.29
Vats et al [1]	52.91	58.45	85.14	89.46
Balaji et al [2]	68.53	73.77	92.50	93.68
Ours	**77.31** ↑8.58	**81.92** ↑8.15	**96.79** ↑4.29	**94.70** ↑1.02

RESULTS

Table III: Results with and without KfID Module. (†) - with the KfID module.

Dataset	Test Acc	Challenge Acc
Ice Hockey	61.71	-
Baseball	88.43	-
SoccerNet	**35.65**	**35.98**
Ice Hockey (†)	96.79 ↑35.08	-
Baseball (†)	94.70 ↑5.73	-
SoccerNet (†)	**77.31** ↑41.66	**81.92** ↑45.94

ABLATION STUDIES

Table VI: Comparison of different Backbones and masking strategies on the SoccerNet dataset.

Backbone	MAE pretraining	Masking Strategy	Test Acc
ResNet-18	✗	-	58.62
ResNet-34	✗	-	61.29
ResNet-152	✗	-	65.10
ViT-B	✓	Zeroing-Out	75.83
ViT-B	✓	Gaussian Blur	76.47
ViT-B	✓	**Motion Blur**	**77.31**

Table V: Impact of feature extractors and metrics for $\mathcal{L}_{\text{siamese}}$ on our overall model performance.

Feature Extractor	ℓ_2-loss	ℓ_1-loss	Cosine Similarity
VGG	**76.30**	76.21	74.52
ResNet	76.45	**77.31**	74.90
InceptionNet	75.84	**75.93**	74.66
AlexNet	74.38	**74.41**	73.93

ABLATION STUDY

Table 5. Ablation study on different heads for the loss function

HO	DW	LC	Test Acc
✓			55.71
✓	✓		62.39
✓	✓	✓	65.14
	✓	✓	63.77
	✓		**68.53**

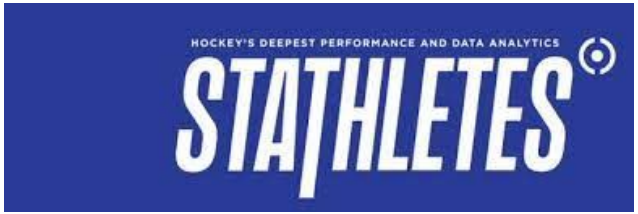
Table 6. Ablation study on different training sequence length

Sequence Length	Test Acc
10	62.82
20	65.45
30	66.52
40	**68.53**
50	67.03
60	65.80

CONCLUSION

- **Efficacy of our *d*-MAE Module**: We demonstrate that incorporating our novel *d*-MAE module results in a performance boost of 12.21% increase on the SoccerNet test set.
- **Significant Improvement on SOTA**: We consistently outperform the existing state-of-the-art by ~8%, ~4% and ~1% on the SoccerNet, Ice Hockey and Baseball datasets respectively, underscoring the impact of motion blur in sports videos.

ACKNOWLEDGEMENT



Thank You!

Open to any questions

WHY HSV?

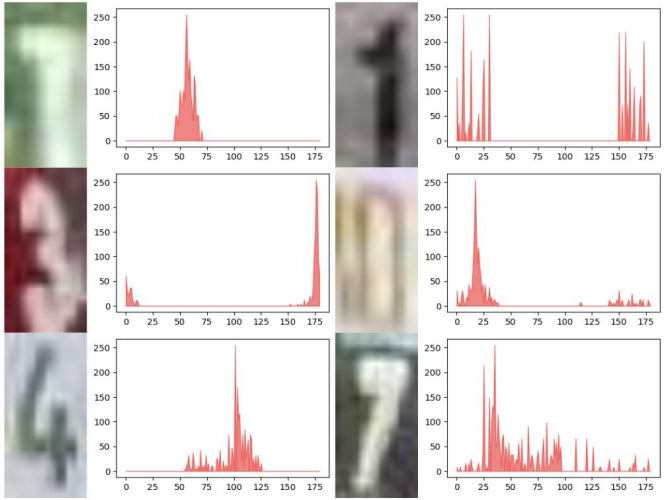


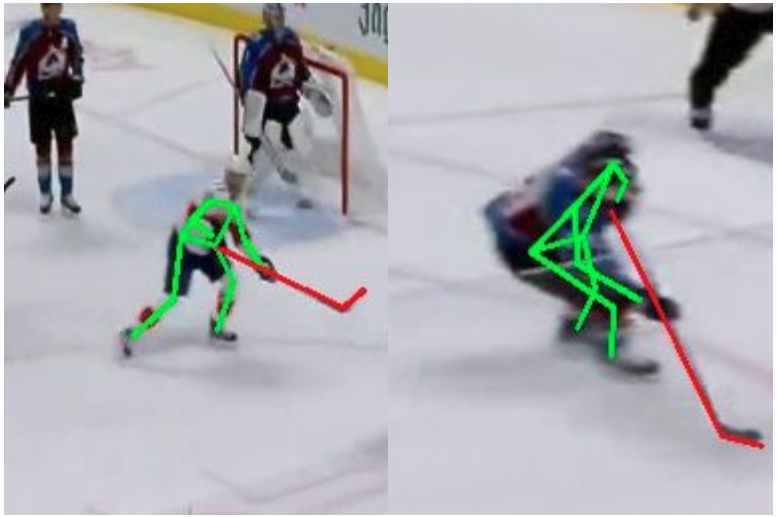
Figure 4. Histogram representation of different jersey's spatial color layout.

INITIAL MODEL

- An ensemble model of 2 MSPN networks
- One trained from scratch on 20 keypoints.
- Another one using a unique transfer learning approach to lift 17 to 20 keypoints.

Qualitative Results

Training from Scratch



Transfer Learning

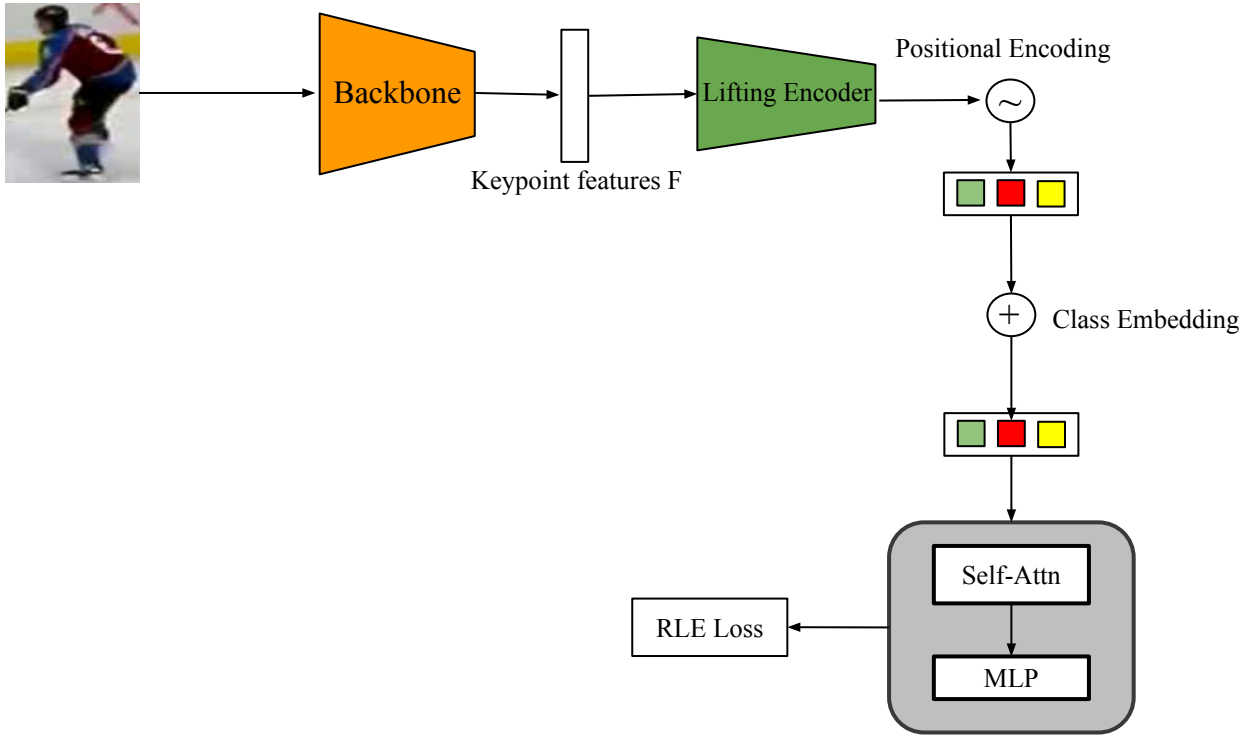
LIMITATIONS

The **presence of multiple players** in one single frame leads to unnecessary information in the image which confuses the input model.

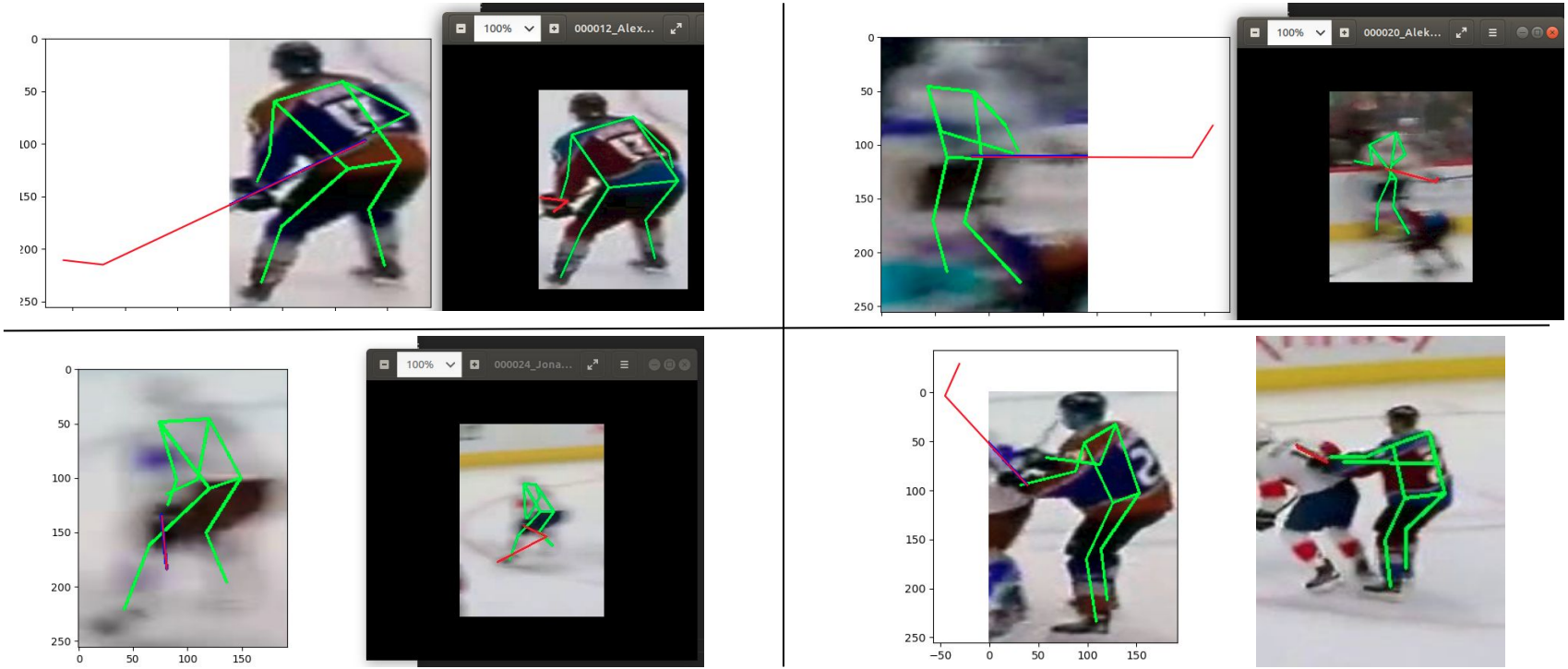
IMPROVED MODEL

- We tried to predict keypoints outside the image.
  - Input image => player bounding box.
  - Output:- player + stick keypoints(not present in input image).
- Now becomes a keypoint regression problem, instead of heatmap regression.

ARCHITECTURE



QUALITATIVE RESULTS



QUANTITATIVE RESULTS

- Comparison of our previous best vs current model

	Ensemble Model	Keypoint Regressor
Training Accuracy	98.69%	99.93%
Validation Accuracy	92.90%	87.38%

ADVANTAGES OF OUR MODEL

- Simple model to predict out-of-image keypoints.
- Can be leveraged for any domain consisting of body extensions, like lacrosse, shoveling, tennis etc.
- Avoids manual decoding of heatmaps and predicts keypoints directly.
- Showcases that the relationship between objects can be exploited to refine each other’s pose.

PLAN OF ACTION

- Creating other datasets for generalizability.
- Exploring other ideas such as using segmentation masks to add more prior to the model.
- Adding other information such as the role of the player(defensemen, forwards etc).