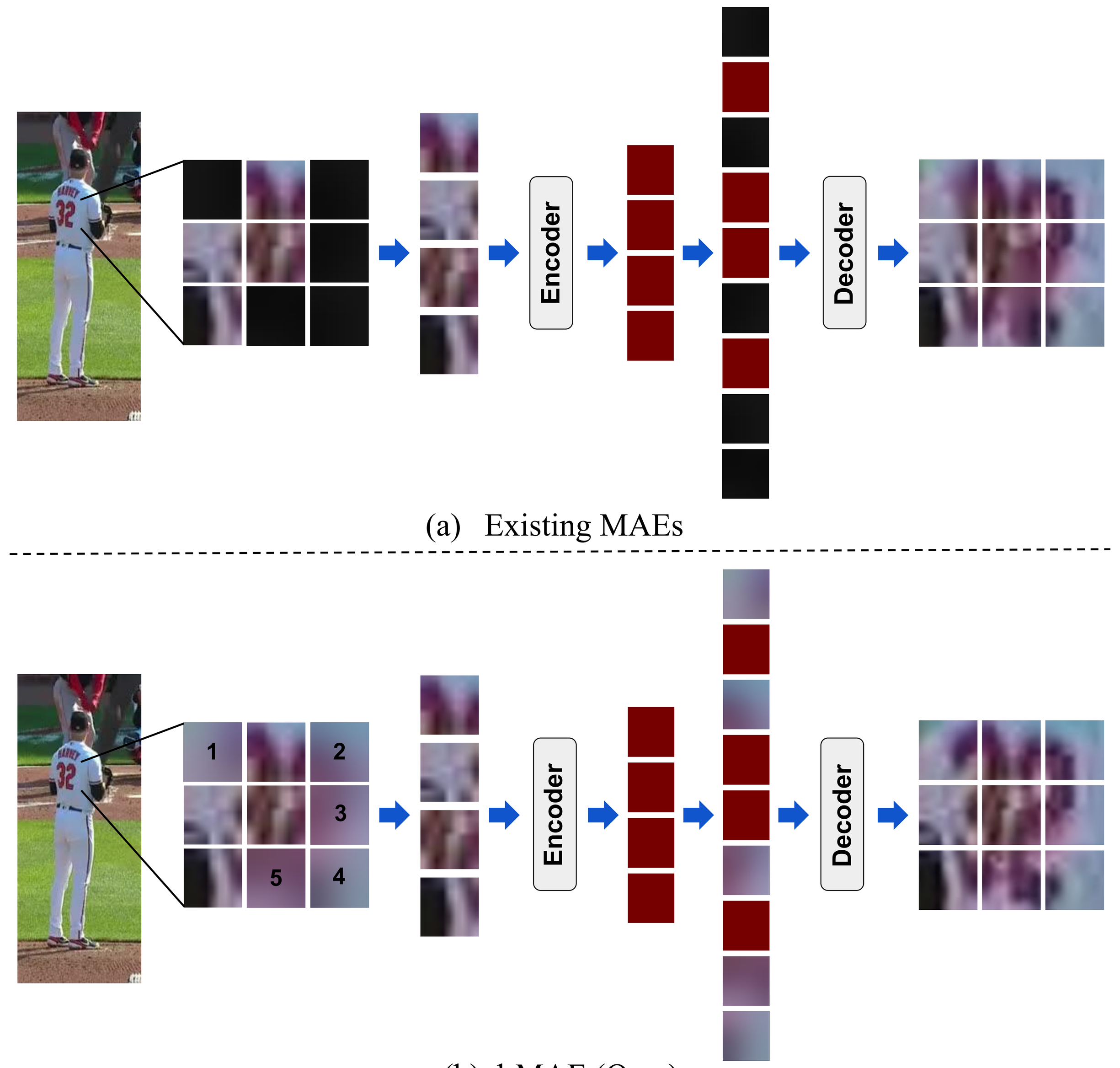
Domain-Guided Masked Autoencoders for Unique Player Identification

Bavesh Balaji, Jerrin Bright, Sirisha Rambhatla, Yuhao Chen, Alexander Wong, John Zelek and David Clausi
University of Waterloo, Waterloo, Ontario, Canada

Key Contributions

- A novel jersey number recognition network that utilizes MAEs coupled with a transformer decoder to capture robust features from low-resolution blurred tracklets.
- A new domain-guided masking strategy, termed *d*-MAE, specifically tailored to player identification, enhancing model robustness to motion blur.
- Refinement of the KfID module [1] by improving its jersey number localization and it's ability to capture fine-grained semantic representations of keyframes.
- Addressing the issue of limited data, we introduce a keyframe fusion technique to augment meaningful data, thereby enriching the training process.
- Validation of our model outperforming SOTA methods on three large-scale datasets spanning different sports.

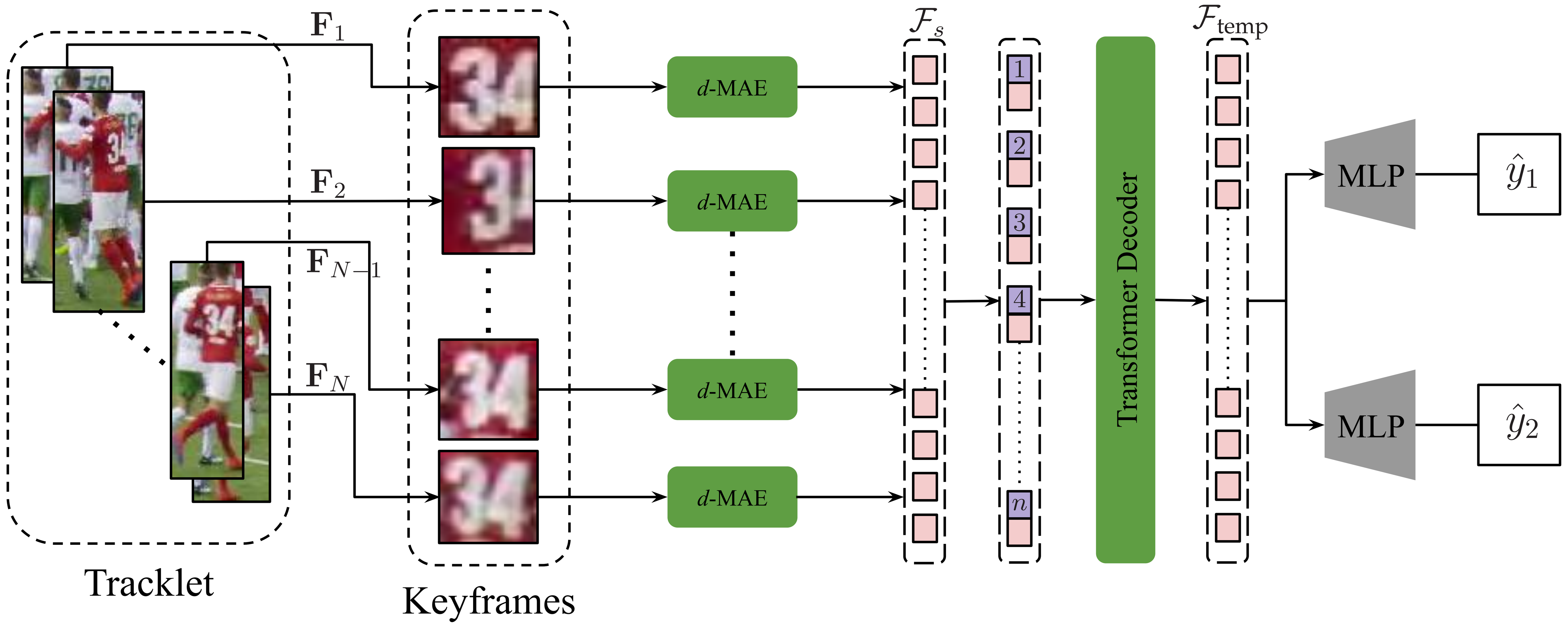
Overview



(a) Existing MAEs

(b) d-MAE (Ours)

Methodology



Overall architecture. Given a tracklet $\mathbb{T}$ consisting of N frames, we pass $\mathbb{T}$ through the KfID module to extract $n \leq N$ keyframes that contain the jersey number. Each keyframe is passed as an input to our *d*-MAE encoder to extract spatial features $\mathcal{F}_s$. These features are then fed to the temporal transformer decoder to extract temporal features $\mathcal{F}_{\text{temp}}$. Two classification heads are utilized to compute the predicted digits of the jersey number $\hat{y}_1$ and $\hat{y}_2$ respectively.

Loss Functions

Siamese Loss:

$$\mathcal{L}_{\text{siamese}} = ||h(\hat{\mathbf{I}}) - h(\mathbf{I})||_1 \quad (1)$$

Multi-head Classification Loss:

$$\mathcal{L}_{\text{class}} = -\sum_{i=0}^{10} y_1^i \log \hat{y}_1^i - \sum_{j=0}^{10} y_2^j \log \hat{y}_2^j \quad (2)$$

Ablation Study

- Comparison with backbones and masking strategies

Backbone	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**

- Comparison of our model with and without KfID

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**

- Impact of feature extractors and metrics for $\mathcal{L}_{\text{siamese}}$

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

Quantitative Results

Method	SoccerNet	Ice Hockey	Baseball
Gerke et al	32.57	61.20	64.47
Vats et al	46.73	83.17	87.61
Li et al	47.85	81.15	88.29
Vats et al	52.91	85.14	89.46
Balaji et al	68.53	92.50	93.68
Ours	**77.31**	**96.79**	**94.70**

Qualitative Results

Performance of our model on two different player tracklets from all three datasets. We find our model's prediction for each image separately and for the entire tracklet (**Pred**). **GT** represents the ground-truth value for the entire tracklet.

Acknowledgement

Paper Preprint:

STATHLETES
NSERC CRSNG
Orioles
Mitacs

References

[1] Bavesh Balaji, Jerrin Bright, Harish Prakash, Yuhao Chen, David A. Clausi, and John Zelek. Jersey number recognition using keyframe identification from low-resolution broadcast videos. In *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, MMSports '23, 2023.