Seeing Beyond the Crop: Using Language Priors for Out-of-Bounding Box Keypoint Prediction

Bavesh Balaji, Jerrin Bright, Sirisha Rambhatla, Yuhao Chen,
John Zelek and David Clausi

University of Waterloo · VIP lab (Vision and Image Processing Research Group, vip.uwaterloo.ca) × Critical ML

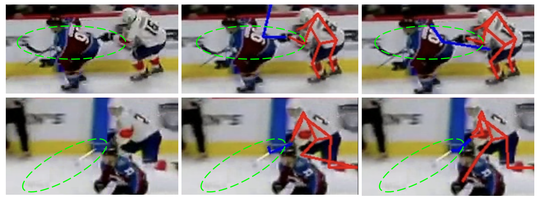
Neural Information Processing Systems

Preliminaries

- **Pose Estimation:** Given a monocular RGB image, we want to predict the 2D coordinates of human joints along with the object that they hold, *aka* extensions.
- **Top-down paradigm:** Isolate each entity by cropping them using bounding boxes, and them perofrm pose estimation individually.

Motivation

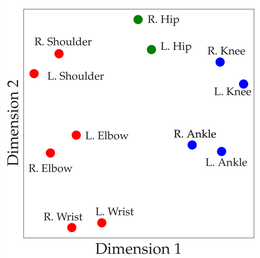
- **Including the extension** in the bounding box leads to the addition of **unnecessary visual features**!
  - Existing methods expect all keypoints of interest to be present in the bounding box.
  - This leads to larger bounding boxes, introducing noisy features that hinder performance.



- Given that the pose of the human and their extension are highly correlated (Yao *et al.* [1]), what if we **don't capture the extension** in the bounding box and still predict it?
- If so, how to inform the model about these *unseen* keypoints?

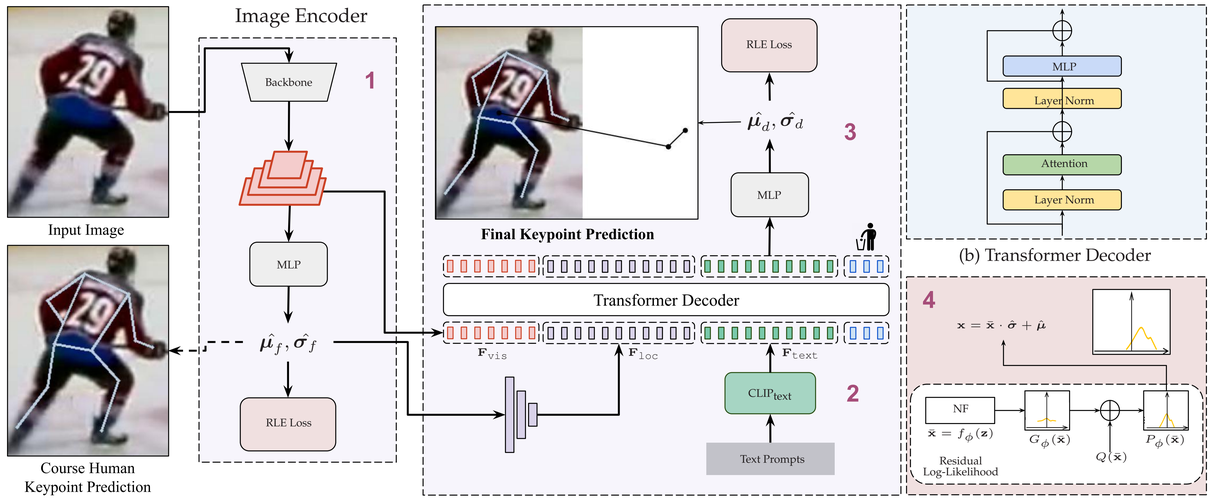
Contributions

- Reformulate the extension pose estimation task as an unseen keypoint prediction problem.
- Utilize keypoint-specific text prompts to induce semantic context of unseen keypoints.



  - Previous works align the learnable image embeddings to the frozen text embeddings.
  - However, text embeddings do not maintain global positional structure.
- Initialize learnable keypoint tokens using these text embeddings and utilize a transformer to find the relationships between different joints.
- Introduce two first-of-its-kind datasets for extension pose estimation collected from real-world ice hockey and lacrosse games respectively.
- Outperform SOTA on our datasets and CrowdPose.

Methodology



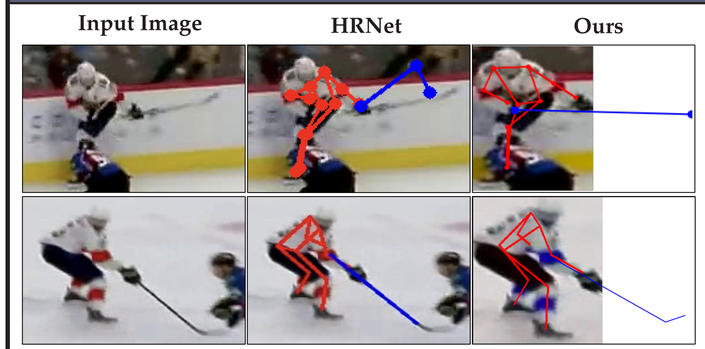
(a) Proposed `TokenCLIPose` Architecture (c) RLE Loss

(1) Initially, the cropped image that excludes the extension is passed through a backbone network to generate dense multi-scale visual features $\mathbf{F}_{\texttt{vis}}$ and coarse human keypoint locations $\mathbf{F}_{\texttt{loc}}$.
(2) Joint-specific text prompts are used to generate text embeddings for each keypoint. Learnable keypoint tokens for each joint $\mathbf{F}_{\texttt{text}}$ are then initialized using these text embeddings.
(3) The image, location and text features are fed to a standard transformer layer which outputs the final predictions.
(4) We use **residual log-likelihood loss** to make our model more robust to noisy keypoints.
  - We formulate regression task as a distribution learning problem and adopt MLE.
  - We leverage normalizing flows to learn how the output deviates from the ground truth and minimize it.

RLE Loss: $\mathcal{L}_{RLE} = -\log P_{\Theta,\phi}(\mathbf{x}|\mathcal{I})\Big|_{\mathbf{x}=\boldsymbol{\mu}_g} = -\log P_\phi(\bar{\boldsymbol{\mu}}_g) + \log \hat{\boldsymbol{\sigma}} = -\log Q(\bar{\boldsymbol{\mu}}_g) - \log G_\phi(\bar{\boldsymbol{\mu}}_g) - \log s + \log \hat{\boldsymbol{\sigma}}.$

Total Loss: $\mathcal{L} = \mathcal{L}^f_{RLE} + \mathcal{L}^d_{RLE}$, where $\mathcal{L}^f_{RLE} = -\log P_{\Theta_f,\Phi_f}(\mathbf{x}|\mathcal{I})\Big|_{\mathbf{x}=\boldsymbol{\mu}_g}$ and $\mathcal{L}^d_{RLE} = -\log P_{\Theta_d,\Phi_d}(\mathbf{x}|\mathcal{I})\Big|_{\mathbf{x}=\boldsymbol{\mu}_g}$

Qualitative Results



(a) Qualitative Comparison on Ice hockey dataset.

(b) Qualitative results on Lacrosse and CrowdPose datasets.

Quantitative Results

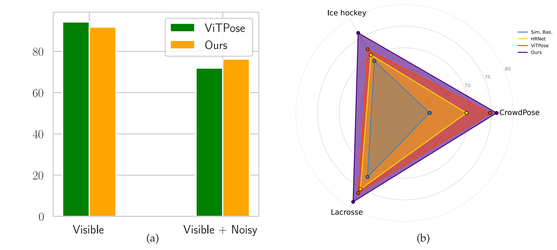
- PCKh@0.5 for joint hockey player and stick pose estimation.

Method	Backbone	Input Resolution	Body	Butt End	Stick Heel	Stick Toe	Mean
SimpleBaseline	ResNet-50	256x192	93.59	69.57	57.19	52.76	68.83
MSPN	-	256x192	93.61	70.30	59.21	55.69	69.70
HR-Net	HRNet-W48	256x192	94.90	71.48	60.29	55.36	70.44
TokenPose-L/D24	HRNet-W48	256x192	95.13	70.96	60.93	56.27	70.82
ViTPose	ViT-B	256x192	95.61	71.94	61.33	58.80	71.92
`TokenCLIPose`	ResNet-50	256x192	95.81	74.86	65.79	65.08	74.92
`TokenCLIPose`	MSPN	256x192	97.17	75.41	66.70	66.34	75.53
`TokenCLIPose`	HRNet-W48	256x192	97.37	75.94	67.82	66.15	76.28
Improvement	-	-	1.76% ↑	4.00% ↑	6.49% ↑	7.35% ↑	4.36% ↑

- Zero-shot results for Lacrosse.

Method	Backbone	Body	Butt End	Stick Heel	Mean
SimpleBaseline	ResNet-50	94.73	67.28	53.99	72.00
MSPN	-	95.84	70.68	57.40	74.64
HR-Net	HRNet-W48	95.92	71.35	58.41	75.22
ViTPose	ViT-B	95.77	72.85	60.18	76.26
`TokenCLIPose`	HRNet-W48	97.24	76.60	65.01	78.61
Improvement	-	1.47% ↑	3.75% ↑	4.83% ↑	2.35% ↑

We achieve 4.36% and 2.35% improvement on our two datasets. Significant improvement on stick joints (4% and 6.49%)!



(a) We significantly improve the prediction of noisy keypoints. **(b)** This also translates to standard occlusion datasets like CrowdPose where we outperform SOTA top-down methods.

Acknowledgement

Paper Preprint:

STATHLETES · NSERC CRSNG · Mitacs

References

[1] Bangpeng Yao and Li Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 17–24, 2010.