Language Guided Out-of-Bounding Box Pose Estimation for Robust Ice Hockey Analysis

by

Bavesh Balaji

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2024

© Bavesh Balaji 2024

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This thesis contains material from two papers for which I was the lead author. As the lead author, I was responsible for the conceptualization, writing code as well as writing and submitting manuscripts.

Chapter 4 of this thesis contains content from the previously published paper: **Bavesh Balaji** and David Clausi.“ Towards Agile Human Pose Estimation: A Case Study in Ice Hockey Analytics. ” In *Journal of Computational Vision and Imaging Systems*, Volume 9, Issue 1, 2023.

Chapter 6 of this thesis contains content from a paper that is under review at *Neural Information Processing Systems (NeurIPS)* 2024: **Bavesh Balaji**, Jerrin Bright, Sirisha Rambhatla, Yuhao Chen, John Zelek and David Clausi. “ Seeing Beyond the Crop: Using Language Priors for Out-of-Bounding Box Keypoint Prediction ”

Abstract

Accurate estimation of human pose and the pose of interacting objects, such as hockey sticks, is fundamental in vision-driven hockey analytics and crucial for tasks like action recognition and player assessment. Estimating 2D keypoints from monocular video is challenging, particularly in fast-paced sports such as ice hockey, where motion blur, occlusions, bulky equipment, color similarities, and constant camera panning complicate accurate pose prediction.

This thesis addresses these challenges with contributions on three fronts. First, recognizing the lack of an existing benchmark, we present a comparative study of four state-of-the-art human pose estimation approaches using a real-world ice hockey dataset. This analysis aims to understand the impact of each model on ice hockey pose estimation and investigate their respective advantages and disadvantages.

Building on insights from this comparative study, we develop an ensemble model for jointly predicting player and stick poses. The ensemble comprises two networks: one trained from scratch to predict all keypoints, and another utilizing a unique transfer learning paradigm to incorporate knowledge from large-scale human pose datasets. Despite achieving promising results, we observe that these top-down approaches yield suboptimal outcomes due to constraints such as requiring all keypoints to be within a bounding box and accommodating only one player per bounding box.

To overcome these issues, we introduce an image and text based multi-modal solution called `TokenCLIPose`, which predicts stick keypoints without encapsulating them within a bounding box. By focusing on capturing only the player in a bounding box and treating their stick as *missing*, our model predicts out-of-bounding box keypoints. To incorporate the context of the *missing* keypoints, we use keypoint-specific text prompts to leverage the rich semantic representations provided by language.

This dissertation's findings advance the state-of-the-art in 2D pose estimation for ice hockey, outperforming existing methods by 2.6% on our dataset, and provide a robust foundation for further developments in vision-driven sports analytics.

Acknowledgements

First, I would like to thank my supervisors Prof. David A. Clausi and Dr. Sirisha Rambhatla for their immense support and encouragement without which this thesis would not have been possible. Thank you for being wonderful mentors and being a source of inspiration.

I would like to thank my committee members Dr. Charbel Azzi and Dr. Elliot Creager for their time and constructive feedback during my comprehensive exam. I would also like to thank Stathletes Inc, for providing me with necessary resources to conduct this research.

A huge thanks to the members and alumni of Vision and Image Processing research group, especially my collaborators Jerrin Bright and Harish Prakash for fruitful collaborations and discussions. I would like to thank my family: my parents Balaji Srinivasan and Subhasri Balaji, my sister Baavana Balaji and my brother-in-law Ashwath Santhanam for providing me with love and support.

Lastly, I would like to thank my friends outside the lab for their help and support in my life, be it work or personal.

Dedication

This thesis is dedicated to my niece Dhwani.

Table of Contents

Author's Declaration ii

Statement of Contributions iii

Abstract iv

Acknowledgements v

Dedication vi

List of Figures xi

List of Tables xiii

1 Introduction 1

1.1 Premise 1

1.2 Motivation 2

1.3 Scope 2

1.4 Challenges 2

1.5 Contributions 3

1.6 Thesis Structure 4

2 Background **5**
2.1 Single-Person Pose Estimation . . . 5
2.1.1 Part-based methods . . . 5
2.1.2 Deep Learning based methods . . . 6
2.2 Multi-Person Pose Estimation . . . 7
2.2.1 Single-stage Methods . . . 7
2.2.2 Two-Stage Methods . . . 8

3 Datasets and Metrics **10**
3.1 Datasets . . . 10
3.1.1 Ice Hockey Dataset . . . 10
3.1.2 Lacrosse Dataset . . . 14
3.1.3 CrowdPose . . . 14
3.2 Evaluation Metrics . . . 15
3.2.1 Object Keypoint Similarity (OKS) . . . 15
3.2.2 Percentage of Correct Keypoints with Head-Normalization . . . 15

4 Benchmarking Human Pose Estimation Techniques for Ice Hockey: A Preliminary Study **17**
4.1 Introduction . . . 17
4.2 2D Pose Estimation . . . 18
4.2.1 MultiStage Pose Network . . . 18
4.2.2 High Resolution Network . . . 19
4.2.3 Distribution-Aware Coordinate Representation of Keypoint . . . 20
4.2.4 TransPose . . . 21
4.3 Comparison Study . . . 21
4.3.1 Dataset . . . 22
4.3.2 Implementation Details . . . 22
4.3.3 Results . . . 22
4.4 Conclusion . . . 25

5 Ensemble Model for Joint Player and Stick Pose Estimation **27**
5.1 Introduction . . . 27
5.2 Methodology . . . 28
5.2.1 Training from Scratch . . . 28
5.2.2 HyperMSPN: Transfer Learning . . . 28
5.2.3 Ensemble Model . . . 28
5.3 Experiments . . . 29
5.3.1 Implementation Details . . . 29
5.3.2 Model Settings . . . 29
5.3.3 Results . . . 29
5.4 Limitations . . . 33

6 Seeing Beyond the Crop: Using Language Priors for Out-of-Bounding Box Keypoint Prediction **34**
6.1 Introduction . . . 35
6.2 Related Work . . . 37
6.3 Method . . . 38
6.3.1 Problem Formulation . . . 38
6.3.2 Network Architecture . . . 38
6.3.3 Loss Function . . . 40
6.4 Experiments . . . 41
6.4.1 Ice Hockey Dataset . . . 41
6.4.2 Lacrosse Dataset . . . 42
6.4.3 CrowdPose Dataset . . . 44
6.4.4 Ablation Studies . . . 45
6.5 Conclusion . . . 48

7 Conclusion and Future Work **50**

7.1 Summary of Contributions 50

7.2 Future Research 51

7.2.1 Action recognition using joint pose 52

7.2.2 Unique player identification using player handedness 52

7.2.3 Finding Tape Color 52

7.2.4 Temporal 2D pose estimation 52

7.3 Thesis Applicability 53

7.4 Thesis Impact 53

References **54**

List of Figures

2.1 Taxonomy of 2D Pose Estimation Methods 6
2.2 A visualization of the heatmap representation 7

3.1 Pictorial Representation of the keypoints used to predict the pose of a hockey player and their stick. Image courtesy: VIP-HARPET dataset [1]. 11
3.2 Ice hockey dataset 12
3.4 Visualization of the distribution of visibility of each keypoint across frames. 13

4.1 Overall Architecture of MSPN [2]. 18
4.2 Overall Architecture of HRNet [3] 19
4.3 Overall Architecture of TransPose [4]. 21
4.4 Inference of the implemented 2D pose estimators without occlusion. 25
4.5 Inference of the implemented 2D pose estimators with occlusion. 25

5.1 Qualitative comparison of the two MSPN models trained on our ice hockey dataset. 30
5.2 Complimentary nature of the errors 32
5.3 Limitations of our ensemble baseline solution 33

6.1 Difference between existing networks and our network 36
6.2 Overall Architecture of `TokenCLIPose` 38
6.3 Detailed overview of Residual Log-Likelihood Loss 40
6.4 Qualitative Comparison of `TokenCLIPose` with the baseline 42

6.5 Qualitative Results of our `TokenCLIPose` on the CrowdPose dataset. 44
6.6 Qualitative results of `TokenCLIPose` for the extension pose estimation task 48

List of Tables

3.1 Keypoints 11

4.1 Per joint and total training accuracy of the fine-tuned networks. 23
4.2 Per joint and total test accuracy of the fine-tuned networks 24

5.1 Quantitative Comparison of all three models 31

6.1 Comparison with SOTA on ice hockey dataset 43
6.2 Zero-shot comparison with SOTA on Lacrosse Dataset 43
6.3 Comparison with SOTA on CrowdPose 45
6.4 Effect of Different Modalities 45
6.5 Impact of different attention mechanisms 46
6.6 Impact of Text Prompts 46
6.7 Impact of Bounding Boxes 47
6.8 Comparison with LAMP on CrowdPose dataset 47

Chapter 1

Introduction

1.1 Premise

Sports analytics involves extracting valuable insights from sports games to analyze and enhance the performance of players and teams. This field has expanded significantly with technological advancements, catering to the massive global sports market. Various sports enjoy widespread popularity, with millions of fans attending events ranging from local matches to international tournaments. Hockey, in particular, boasts a substantial following, especially in North America, where the NHL generated 2.27 billion USD in revenue during the 2022/23 season [1]. Teams constantly seek to outshine their competitors, and sports analytics plays a crucial role by providing detailed statistics about players and teams.

Starting from the 2021 season, the NHL has introduced the puck and player tracking technologies. This includes the installation of sensors on the puck and player jerseys [2]. Furthermore, the sensors require the installation of 14-18 infrared cameras to capture the signals emitted by these sensors, which are then utilized to track the players and the puck in three dimensions. While this method is accurate, it incurs significant equipment and personnel costs due to installation of the cameras and maintaining sensors on the puck and players.

To overcome these challenges and provide non-invasive and affordable methods for analyzing a hockey game, video analytics are employed. This involves the computer-assisted analysis of images and videos to extract meaningful game and player data. Given the

[1] https://www.statista.com/statistics/764234/nhl-gate-receipts-ticketing-revenue/
[2] https://www.espn.com/nhl/story/_/id/38724195/nhl-brings-advanced-puck-tracking-stats-public

vast amount of data generated during games, manual processing is impractical, making advancements in computer vision essential for automation. These techniques enable the extraction of information that would otherwise be nearly impossible to obtain due to the game's fast pace, multiple participants, and data volume, facilitating further analysis and insights.

1.2 Motivation

One of the most important aspects of hockey analytics is player assessment and game action recognition. The human motion of every player along with their hockey sticks must be analyzed to effectively evaluate players and recognize actions. A key tool for this analysis is 2D human pose estimation, which focuses on localizing anatomical keypoints of the human body. However, the dynamic nature of the game presents significant challenges, making it difficult to rely solely on body posture for accurate assessments. Thus, relying solely on body posture might not always be sufficient. Objects with which people interact (tools, equipment, or sports gear) can hold significant information about the activity and anatomy of the person. Capturing the pose of these objects, in our case, the hockey stick, enriches the analysis of human motion and also aids in refining the pose of the person by understanding the pose of the stick [1]. Developing a robust automated solution for predicting the 2D pose of players along with their sticks is highly impactful as it can influence the decisions of scouts, coaching staff and team management. Such a solution mitigates the need for manual annotations to study the pose of a player. An automated solution would also provide an easier, more universally accessible approach to assess any players for scouting and recruiting with the only need being a game video from a suitably placed camera in the rink.

1.3 Scope

This thesis primarily focuses on automated 2D pose estimation of players and their hockey sticks from broadcast video feeds of hockey games.

1.4 Challenges

The challenges associated with pose estimation in ice hockey include the following items:

Motion Blur. The motion of players in hockey is is extremely fast compared to the motion of players in other team sports such as basketball and soccer. There is also significant camera movement primarily due to panning when players move from one offensive zone to the other. Significant motion blur with standard broadcast rates and resolution make it difficult to interpret the pose of the player and extremely difficult to identify their sticks as part of the pose.

Occlusions and Bulky equipment. Compared to other sports such as soccer and basketball, accurately identifying the pose of players is much more challenging in ice hockey due to the players wearing bulky equipment [1, 5, 6] and helmets that largely occlude body joints, especially reducing the visibility of head and hand keypoints significantly. Furthermore, the crowded nature of the game coupled with events such as face-offs lead to player-player occlusions and self-occlusions making detection of player and stick keypoints very challenging.

Lack of Publicly Available Datasets. Pose estimation for ice hockey is in itself highly underaddressed. Available datasets [1, 5] have images captured in contrived training environments that do not emulate real ice hockey games. As a result, these datasets do not capture the non-idealties present in real-world hockey data.

1.5 Contributions

1. There has been no published work that extensively evaluates and analyzes state-of-the-art (SOTA) pose networks on ice hockey datasets. Therefore, we compare and contrast the SOTA pose networks and analyze their performance and failure modes when applied to ice hockey game video.

2. Existing research on ice hockey player and stick pose estimation [1] utilize older techniques such as Stacked Hourglass networks [7]. However, recent advancements in 2D pose estimation have led to powerful solutions [2, 3, 4] that outperform hourglass networks. Hence, we develop a competitive baseline solution using MSPN [2] for ice hockey pose estimation by using the knowledge gained from benchmarking the SOTA pose estimators.

3. To overcome the limitations of our MSPN-based solution, we introduce a novel solution to reliably predict stick keypoints from player keypoints using language priors and introduce the concept of *out-of-bounding box keypoint prediction*. We re-consider these *out-of-bounding box* keypoints as *missing/invisible* keypoints and approach it from the

point-of-view of occlusions. We further demonstrate the effectiveness of our model in overcoming occlusions by conducting rigorous experiments on the CrowdPose dataset, the largest open source dataset for occluded 2D human pose estimation.

4. Realizng the lack of real-world ice hockey datasets that emulate game situations, we create an Ice Hockey dataset from broadcast NHL videos. To showcase the generalization capabilities of our model, we also curate a real-world lacrosse dataset from a YouTube video and demonstrate our model's robustness to different domains.

1.6 Thesis Structure

This thesis is organized into seven chapters. Chapter 2 presents a literature review of the existing publications, datasets, and metrics on 2D human pose estimation and their applications in sports. Following the literature review, Chapter 3 introduces the ice hockey and lacrosse datasets and provides detailed descriptions of the various challenges present in the data. Chapter 4 compares and contrasts several SOTA pose estimators and analyzes their performance and failure modes in ice hockey. Using the insights gained from Chapter 4, Chapter 5 introduces an ensemble solution for ice hockey, implemented by combining recent pose estimation networks. However, deeper analysis revealed that there are certain limitations in this solution. To overcome these shortcomings, Chapter 6 introduces our novel multimodal solution for predicting out-of-bounding box keypoints using keypoint-specific language priors. Finally, we conclude our thesis in Chapter 7 by providing a summary of the contributions and discussing the limitations and potential for future research.

Chapter 2

Background

In this chapter, an extensive overview of the existing 2D human pose estimation models is provided. 2D human pose estimation can broadly be classified into two categories: single-person pose estimation and multi-person pose estimation. Single-person pose estimation further encompasses traditional approaches that use part-based modelling of humans and deep-learning based approaches. Multi-person pose estimation can further be grouped into two categories: single-stage and two-stage methods. The taxonomy of existing works is illustrated in Fig. 2.1. This chapter provides a detailed outline of the existing works on single-person and multi-person pose estimation in Sections 2.1 and 2.2 respectively.

2.1 Single-Person Pose Estimation

2.1.1 Part-based methods

Human pose estimation is a fundamental problem in computer vision that has been researched for a long time. It aims at predicting human anatomical keypoints from independent images. The first work dates back to 1971, where Fischler and Elschlager [8] introduced a new representation called pictorial structures to represent objects in images. Contrary to the earlier methods of line type drawings, they devised a deformable part-based method where each part referred to an individual segment of body such as head or torso. After this work, all works for pose estimation used part-based models along with tree-based probabilistic graphical approaches to model the spatial relationships. Felzenszwalb and Huttenlocher [9] developed an efficient algorithm for global matching of pictorial structures,

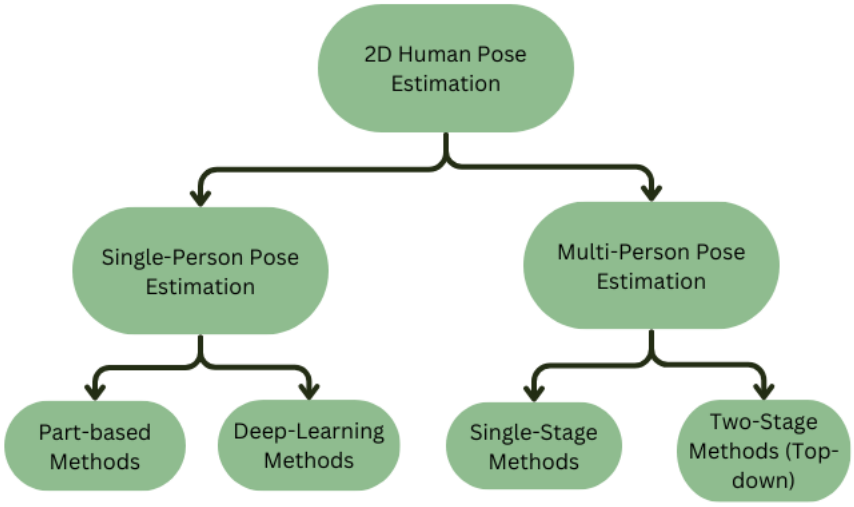


Figure 2.1: Taxonomy of 2D Pose Estimation Methods

which they used to estimate articulated human poses. Their approach employed a geometric model of the human body composed of object parts. With the advent of machine learning methods, later works sought to leverage them for detecting body parts and modelling their connections. Ronfard *et al.* [10] used support vector machines (SVMs) to classify image patches; Andriluka *et al.* [11] use shape context descriptors to extract local features that best represent the different parts for part-based pose estimation; and Pischulin *et al.* [12] used histogram of gradients (HOGs) [13] to condition pictorial structures. However, these models could not model the spatial location effectively and handle occlusions. Hence, with the advent of convolutional neural networks (CNNs), a multitude of works has been conducted on the use of deep learning approaches for pose estimation as CNNs extract robust spatial features due to their translational equivariance property.

2.1.2 Deep Learning based methods

Toshev et al. [14] initially formulated pose estimation as a regression problem of finding body joints, and used CNNs to estimate the poses. To improve the precision of the model without increasing the input resolution, they use a cascade of regressors that iteratively refine the initial pose predicted by the regressor at stage 1. Tompson *et al.* [15] argued that directly regressing 2D coordinates of keypoints is a non-linear mapping and highly difficult to learn. Hence, Tompson *et al.* [15] converted the 2D coordinates of each joint to keypoint

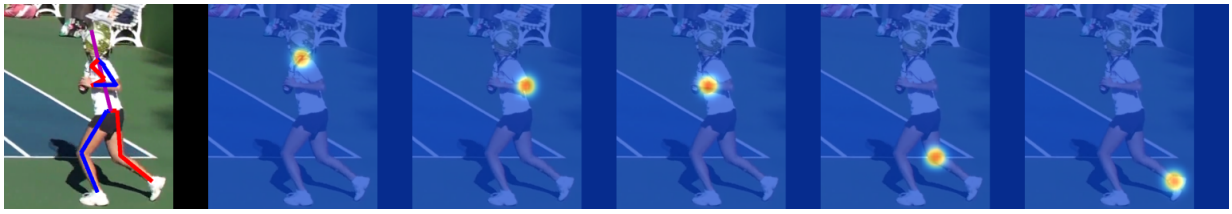


Figure 2.2: A visualization of the heatmap representation depicted in Newell *et al.* [7]
.

heatmaps. Specifically, they converted each joint coordinate to 2D Gaussians centered on that joint location with small variance. Fig. 2.2 depicts the heatmap representation for a sample image. Following this heatmap approach, Newell et al. [7] was the first to use a multi-stage encoder-decoder architecture where each stage consisted of repeated down and upsampling layers with skip connections [16] to extract as much information as possible. Similar to the U-Net architecture [17], several residual sub-modules with numerous skip connections were stacked on top of one another. Additionally, this research publication also introduced intermediate supervision to iteratively refine the heatmaps produced in each stage. Subsequently, many architectures [2, 3, 18, 19, 20, 21, 4, 22] use an encoder-decoder architecture.

2.2 Multi-Person Pose Estimation

2.2.1 Single-stage Methods

Single-stage multi-person pose estimation methods focus on predicting all the human poses in an image in a single forward pass. Pischulin *et al.* [23] proposed the first bottom-up pose estimator that jointly solves object detection and pose estimation. This method partitions an initial set of mutually consistent body paerts and then performs a bottom-up grouping of these body parts by using an integer linear program. However, this method had a very slow inference time. Insafutdinov *et al.* [24] proposed a faster solution DeeperCut by using stronger conditional part-detectors and a newer optimization method. After this work, Cao *et al.* [25] developed OpenPose, an extension of Convolutional Pose Machines [26] that predicts part affinity fields, a set of 2D vector fields that represent the location and orientation of limbs over the image domain. Around the same time, Newell *et al.* [27] extended stacked hourglass networks to predict multiple poses using associative embedding tags. The joints containing the same embedding tag are joined together to form a human

pose. These tags are trained using a contrastive objective where tags of similar poses are brought closer to each other while tags of different poses are pulled away from each other.

Cheng *et al.* [28] leveraged HRNet [3] and added associative embeddings to create the first bottom-up method that maintains high-resolution feature maps throughout the network. Luo *et al.* overcame the scale variation of humans across images and reinvented heatmap regression by adaptively adjusting the standard deviation of each keypoint. Braso *et al.* [29] further built on top of HigherHRNet by utilizing a transformer to group the identity-agnostic keypoints to person centers.

Unlike grouping the detected keypoints in a bottom-up manner, other methods have unified object detection and pose estimation. He *et al.* [30] used the Mask-RCNN segmentation model for pose estimation. Building upon the anchor-free FCOS detector [31], Tian *et al.* [32] directly regressed instance-level keypoints by considering the keypoints as special bounding boxes. Mao *et al.* [33] improved upon this work by incorporating dynamic keypoint heads that are conditioned on human instances. By doing so, they eliminate the need for bottom-up grouping and allow each keypoint head to attend to each human instance respectively. McNally *et al.* [34] extended the representation provided by Tian *et al.* [32] by considering another set of objects called pose objects. These pose objects consists of a bounding box of class "person" and the 2D keypoints of that person. They showcased that using pose objects along with keypoint objects and treating them as a multi-object detection problem yields good results. Yang *et al.* [35] leveraged the same representation and utilized deformable transformers to predict human feature maps and predict keypoint features for each human instance map.

2.2.2 Two-Stage Methods

Two-stage pose estimation methods (a.k.a top-down methods) perform object detection to first isolate all the humans and then perform single-person pose estimation on these isolated entities. While single-stage pose estimators are more efficient and simple especially when there are numerous people in the image, they are still less accurate than two-stage methods. Hence, we focus on two-stage or top-down pose estimation.

Top-down approaches in 2D pose estimation can be broadly classified into two categories: *Heatmap-based pose estimation* has been the *de-facto* standard approach since the introduction of stacked hourglass networks [7]. As mentioned previously, these methods represent discrete (x, y) coordinates as continuous heatmaps where each pixel indicates the likelihood of a specific joint being located at that position. Most methods [2, 3, 7, 36] relied on powerful convolutional networks to extract high-level multi-scale feature maps. While

most research focused on network architectures, a few works investigated the coordinate representation and the heatmap encoding and decoding process [37, 38]. Recently, researchers began exploring transformer-based architectures for pose estimation [18, 4, 21, 22]. Xu *et al.* [18] adopted the original vision transformer [39] and built baselines for pose estimation, showcasing the efficacy of vision transformers. Methods including TransPose and TokenPose [4, 21] used CNNs as feature extractors and utilized transformers to model the relationship between different scale features and the keypoint features respectively.

Regression-based pose estimation, in contrast to the dominance of heatmap-based approaches offers an alternative paradigm. Sun *et al.* [40] leveraged a convolutional backbone to extract feature maps and then utilized an integral operation to directly regress keypoint coordinates. Li *et al.* [41] developed a novel pose regression model aimed at minimizing the distance between the predicted and underlying distribution. However, the global pooling operation used in [41] resulted in a loss of spatial information crucial for reliable pose estimation. More recent works [42, 20, 43] included transformer-based architectures: [42, 20] used an encoder-decoder strategy and were based on DETR [44] and Deformable-DETR [45] respectively.

While heatmap-based methods have achieved high accuracy, they have some limitations:

1. These methods have a non-differentiable post-processing step where they use the *max* operation to extract the final 2D coordinates from these heatmaps.
2. Heatmap representation often leads to quantization errors as the heatmaps are in created in a smaller dimension than that of the original image.
3. They are not designed to predict out-of-bounding-box keypoints.

Therefore, we adopt the regression-based approach for directly estimating the coordinates of all joints in our novel solution presented in Chapter 6.

Chapter 3

Datasets and Metrics

In this chapter, we provide detailed descriptions of the different datasets and evaluation metrics used in this thesis. First, we introduce two first-of-its kind datasets for 2D pose estimation from ice hockey and lacrosse videos respectively. Second, we utilize the CrowdPose dataset to demonstrate our model's capabilities on overcoming occlusions.

3.1 Datasets

3.1.1 Ice Hockey Dataset

The ice hockey dataset, captured from real-world National Hockey League (NHL) videos, presents a challenging environment for pose estimation due to the inherent fast pace of the sport. Players' rapid movements often result in motion blur within frames, making it difficult to distinguish body parts. Furthermore, the nature of the game leads to frequent occlusions, especially when players obstruct each other's bodies. Adding to the complexity is the bulky equipment worn by hockey players, which can obscure keypoints. These combined challenges- motion blur, occlusion, and bulky equipment- make the ice hockey dataset a valuable resource for evaluating the robustness of the pose estimation task.

Our dataset consists of 9 video clips (30-45 seconds each, sampled at 30 fps), each from one broadcast NHL video. The data annotation process was carried out in collaboration with Stathletes, a Canadian company that specializes in the capture and deep performance analysis of ice hockey game data. We utilize the open source Computer Vision Annotation

Tool (CVAT)[1] to annotate the keypoints for multiple players and their hockey sticks in each frame. We followed the standard COCO format to annotate 17 keypoints for the human pose, out of which 5 keypoints correspond to the head. Since the head of each player is covered by a helmet obscuring the head keypoints, only the remaining 12 keypoints are captured. Additionally, we annotate 3 keypoints for the hockey stick (butt end, heel, and toe). All the keypoints identified for pose estimation are listed in Table 3.1. Fig. 3.1 depicts a pictorial representation of each keypoint. Finally, for each annotated frame, we create two versions of the input image: one cropped to include only the player (*standard bounding box*) and another encapsulating both the player and their hockey stick (*extended bounding box*). In total, we generate $11.66K$ pose annotations of approximately humdred different players, with $9.13K$ images from eight clips used for training/validation and $2.53K$ images from the ninth clip used for testing our model's performance. The exact split of training, validation and test sets in terms of video ids is mentioned in the caption of Fig. 3.3. Some of the frames obtained from the broadcast video of the hockey dataset can be visualized in Fig. 3.2.



Figure 3.1: Pictorial Representation of the keypoints used to predict the pose of a hockey player and their stick. Image courtesy: VIP-HARPET dataset [1].

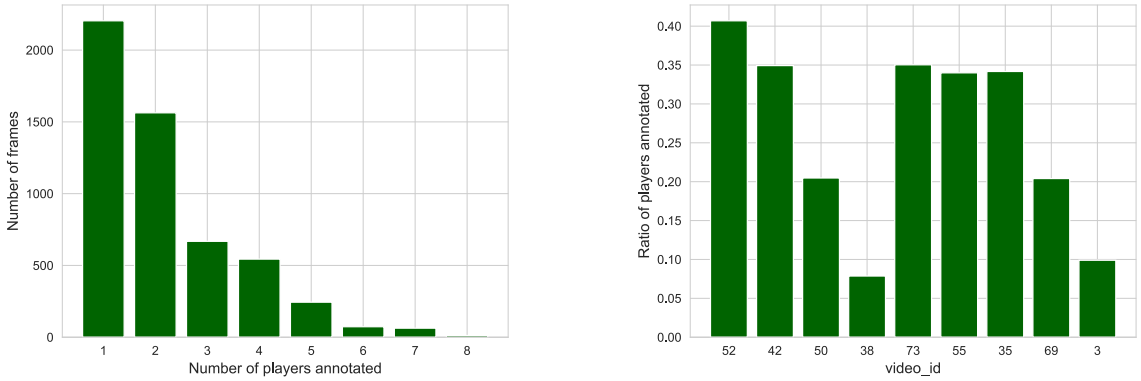
Index	Keypoint
1	Right Shoulder
2	Left Shoulder
3	Right Elbow
4	Left Elbow
5	Right Wrist
6	Left Wrist
7	Right Hip
8	Left Hip
9	Right Knee
10	Left Knee
11	Right Ankle
12	Left Ankle
13	Stick Butt End
14	Stick Heel
15	Stick Toe

Table 3.1: Keypoints

[1]Found online at: https://github.com/openvinotoolkit/cvat



Figure 3.2: Ice hockey dataset



(a) (b)

Figure 3.3: Visualization of the (a) distribution of player annotations per frame, and (b) the number of annotated players per video. Here, video id `55` is used for testing, video id `73` is used for validation while all the other videos are used for training.

Challenges

One of the main challenges of this dataset is the lack of visibility of all keypoints in all the frames. Due to the high degree of motion blur and occlusions, there is a large variance in the visibility of keypoints across frames. This is showcased in Fig. 3.4.

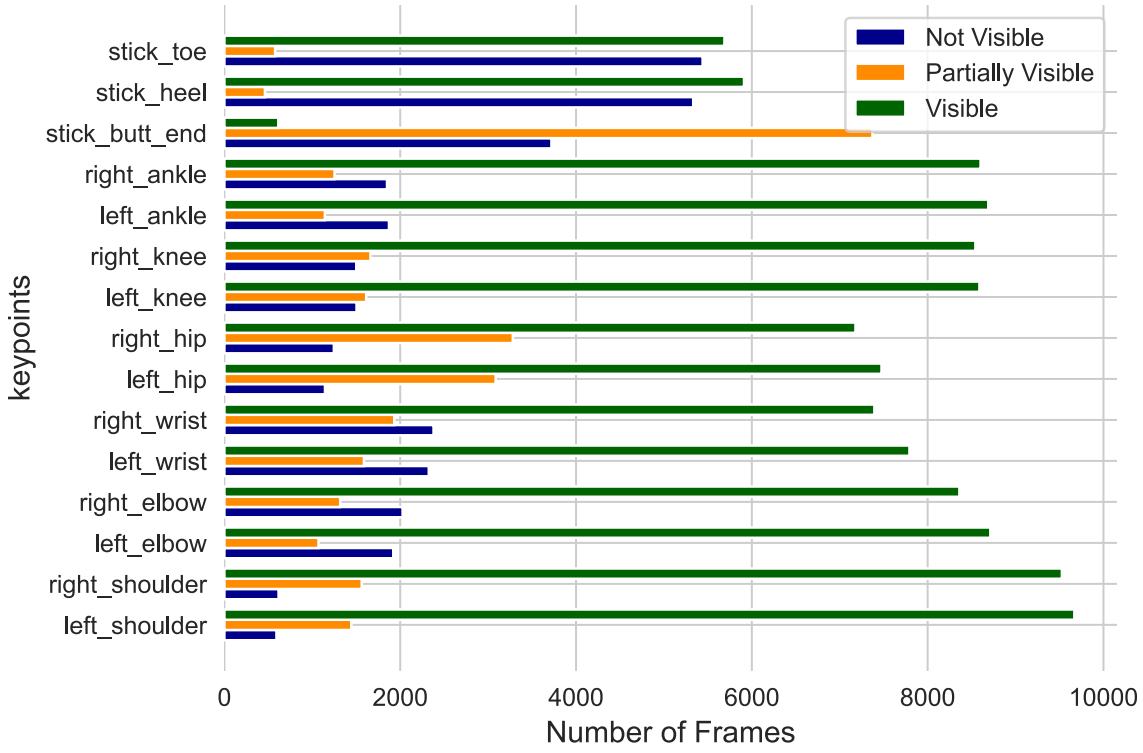


Figure 3.4: Visualization of the distribution of visibility of each keypoint across frames.

From this visualization, it is evident that the stick keypoints are visible only in approximately 50% of frames. This is due to occlusions of the stick caused by players and the boards, and the fast pace of the game, rendering thin and small objects like the hockey stick and puck invisible. Consequently, the paucity of clearly visible data for these keypoints could affect the model's performance in robustly predicting these keypoints.

Another major challenge is the number of annotated players in a frame. Since the video feed that we leverage is from a panning broadcast camera, the field-of-view of the camera is limited to a certain number of players making it difficult to capture the annotations of all 10 players on the rink. Moreover, even when all players are visible within a frame, the process of manually annotating each player can be labor-intensive and time-consuming. Hence, this results in fewer annotations than the number of players present in that frame. This is showcased in Figures 3.3(b) and 3.3(a). These plots show that we typically capture

only 1-2 players per frame, despite there being an average of five players per frame.

3.1.2 Lacrosse Dataset

To study the efficacy of `TokenCLIPose`'s generalization capabilities in predicting the pose of extensions, we curate another small-scale Lacrosse dataset. Similar to ice hockey, it is characterized by motion blur and occlusions, but there are domain differences between the 2 datasets. This dataset consists of 300 pose annotations from one video sampled at 30 fps.

3.1.3 CrowdPose

The CrowdPose dataset is a popular large-scale benchmark dataset for crowded human pose estimation. This dataset is especially difficult for top-down multi-person pose estimators where object detection is used to isolate humans before estimating their pose. This dataset is created using crowded images from three benchmark human pose estimation datasets: COCO [46], AI Challenger [47] and MPII [11]. To determine the amount of crowdedness in an image, a unique matric called Crowd Index is introduced. Specifically, within the bounding box of a particular human, it is defined as the ratio of the number of other humans' keypoints and the number of that human's keypoints. The mathematical equation is as follows:

$$Crowd\ Index = \frac{1}{n} \sum \frac{N_i^b}{N_i^a}, \tag{3.1}$$

where n is the total number of humans in an image, N_i^a is the number of keypoints of a human within their bounding box and N_i^b is the number of keypoints of other humans' within that particular bounding box. By using the equation, they create a dataset with a uniform distribution of Crowd Index contrary to other datasets, making it more difficult for top-down multi-person pose estimators to accurately estimate the keypoints on this dataset.

This dataset contains $12K$ images and $43.4K$ labeled people in their `trainval` set, and $8K$ images with $29K$ labeled people in the `test` set. Following [48, 49, 50], we use the `trainval` set for training and `test` set for evaluation.

3.2 Evaluation Metrics

3.2.1 Object Keypoint Similarity (OKS)

The metrics used for CrowdPose are primarily based on a keypoint similarity measure called Object Keypoint Similarity (OKS) denoted by $m_{OKS} \in \mathbb{R}$. It is a summation of Gaussian-scaled Euclidean distances between predicted and ground-truth keypoints. The mathematical definition is as follows:

$$m_{OKS} = \frac{\sum_i \exp(-\hat{d}_i^2/2s^2k_i^2)\delta(v_i > 0)}{\sum_i \delta(v_i > 0)}, \tag{3.2}$$

where $\hat{d}_i$ is the Euclidean distance between the i-th predicted keypoint coordinate and the corresponding groundtruth, v_i is the visibility flag of each keypoint, k_i is the keypoint-specific standard deviation, and s is the area of the person bounding box which is used to scale k_i. $\delta(.)$ represents the indicator function that returns 1 if the $v_i > 0$ and returns 0 otherwise.

This similarity function along with a threshold is used to predict true positives (tp), false positives (fp) and false negatives (fn). A detected keypoint is considered to be a true positive (tp) if its OKS with a ground-truth keypoint is greater than some threshold β. If the IoU is less than β, the detected object is considered to be a false positive (fp). The number of false negatives (fn) equals the number of ground-truth objects minus the number of true positives. The precision (p) and recall (r) metrics may then computed as follows:

$$p = \frac{tp}{tp + fp} \quad r = \frac{tp}{tp + fn} \tag{3.3}$$

The primary metric used is average precision (AP) which is calculated by averaging the precision over a ten threshold values : $\beta \in [0.5, 0.55, ..., 0.95]$. AP_{50} and AP_{75} indicate the average precision at 0.5 and 0.75 thresholds respectively. Apart from this, AP_E, AP_M and AP_H corresponding to easy, medium and hard images (based on their Crowd Index value) are also computed.

3.2.2 Percentage of Correct Keypoints with Head-Normalization

We evaluate different pose estimation models on our ice hockey and Lacrosse datasets using the percentage of correct keypoints with head-normalization (PCKh) metric [11]. This

metric reports the percentage of keypoint detections that fall within a normalized distance of the ground truth. In general, this metric is determined by computing the length of the head of the person. However, since we do not consider the head keypoints, we use the torso diameter as our normalization factor. Essentially, we find the Euclidean distance between the left shoulder and right hip joints and use them as a threshold to determine correct keypoints. Specifically, keypoints whose Euclidean distance between ground-truth and predicted coordinates are less than or equal the threshold are considered as correct keypoints.

$$PCKh = \frac{\text{number of correct keypoints}}{\text{total number of keypoints}} * 100 \tag{3.4}$$

Chapter 4

Benchmarking Human Pose Estimation Techniques for Ice Hockey: A Preliminary Study

4.1 Introduction

2D pose estimation is a fundamental task in vision-driven hockey analytics, enabling various downstream tasks such as action recognition and player evaluation. However, unlike other settings, hockey poses unique challenges such as occlusion and motion blur, due to its highly dynamic nature. Hence, existing benchmarks do not precisely capture the efficacy of existing methods for ice hockey. In our work, we try to understand the accuracy of existing networks for ice hockey pose estimation by comparing four SOTA models on our ice hockey dataset. We opt for models that follow the top-down approach as it has proven to be the more accurate [51] and dominant [18, 3, 2] paradigm.

The chapter is organized as follows: the methodology of the four pose estimators that we consider are explained in Section 4.2. The implementation details and the results of our study are then explained in Section 4.3.

4.2 2D Pose Estimation

4.2.1 MultiStage Pose Network

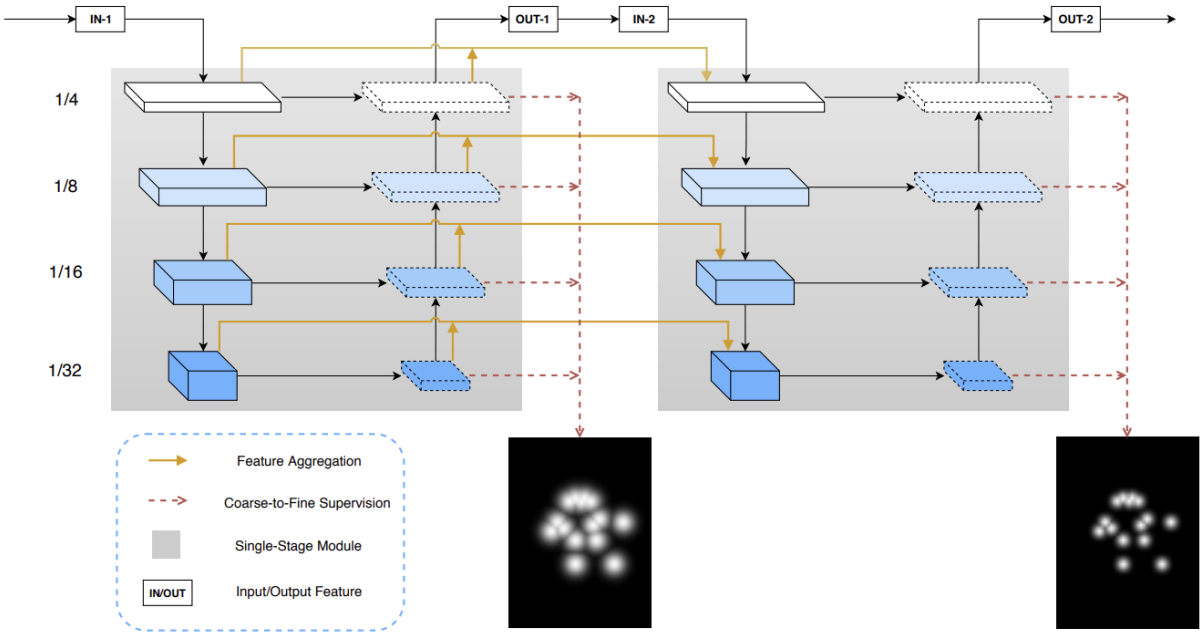


Figure 4.1: Overall Architecture of MSPN [2].

Multi-Stage Pose Networks [2] adopt the top-down approach in two steps. In the first step, manual annotations of all the players on the rink are used to crop the input frames and create multiple images consisting of a single person. The pose estimation network then uses repeated down and up-sampling to continuously refine the estimation of poses. This network mainly proposes three major design improvements on other multi-stage networks. The overall architecture is depicted in Fig. 4.1.

The first one is the equal channel width design followed in all the networks. All the existing networks use the same number of channels in each level of a downsampling module. However, this reduces the size of the feature map as we go down a single stage, making the capture of relevant information more difficult. To solve this problem, this network doubles the number of channels (convolutional kernels) at every level of a downsampling module, maintaining the size of the feature map throughout the stage. This helps the model capture more information in the downsampling module, resulting in better localization of keypoints.

The second improvement made by this architecture is the cross-stage feature aggregation. This network enables us to propagate the features extracted during the initial stages by aggregating them with the features in successive stages. This helps in retaining a lot of information without adding a lot of layers, making the model more robust and foolproof.

The third and most important improvement made by the model is the coarse-to-fine supervision. At the end of every stage, the outputs are converted into Gaussian heatmaps and compared with ground truth heatmaps to refine the localization accuracy. In this network, they perform this intermediate supervision at every stage using decreasing Gaussian kernel sizes (instead of using same size kernels at every stage) as this gives a more accurate estimate of the features extracted.

4.2.2 High Resolution Network

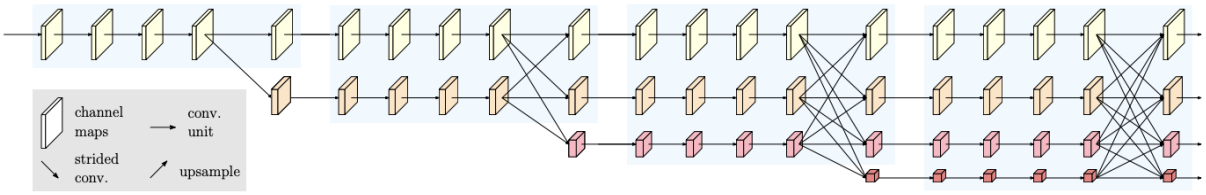


Figure 4.2: Overall Architecture of HRNet [3]

High Resolution Network [3] is a multi-stage pose estimation network that focuses on producing and maintaining accurate high-resolution relationships. This network starts by extracting features at a higher resolution and then goes on reducing the resolution as we go deeper into the architecture. The one unique aspect of this network is the parallel connections across different resolutions, in comparison to the serial connections that are used in other pose estimation architectures. These parallel connections help in effectively maintaining the high-resolution features extracted at the start of the network without losing important information. The other differentiating factor of this model from other existing models is the repeated multi-scale fusion across different resolutions of a single stage. This novel technique aggregates features from different resolutions by either upsampling (using nearest neighbor interpolation followed by 1x1 convolutions) or downsampling (using 3x3 strided convolutions). This concatenation of features within a single stage helps in producing refined and robust representations. The overall architecture is illustrated in Fig. 4.2.

4.2.3 Distribution-Aware Coordinate Representation of Keypoint

The SOTA pose estimation models do not directly take the 2D coordinates of each joint as their input and predict 2D coordinates for every keypoint in a given image. This is mainly because the 2D coordinates do not contain any spatial and contextual information, making pose estimation extremely challenging. Hence, all the existing networks convert these 2D coordinates to heatmaps using a Gaussian kernel to gain some much-needed spatial information. This research [38] focuses on improving this encoding and subsequent coordinate decoding part (after the predictions made by the network) and provides a more principled distribution-aware method.

The standard coordinate encoding methods downsample bounding boxes to a smaller dimension, transforming the ground-truth coordinates accordingly. This downsampling is defined as shown in equation 4.1.

$$g' = (u', v') = \frac{g}{\lambda} = (\frac{u}{\lambda}, \frac{v}{\lambda}) \tag{4.1}$$

In 4.1, $g = (u, v)$ as the ground-truth coordinate and λ is the downsampling ratio. After performing this downsampling, standard methods generally quantize these coordinates using the floor or the ceil function to facilitate kernel generation. However, this causes an inaccurate and biased representation because of quantization error. This implementation solves that problem by eliminating the quantization step and placing the center of the heatmap at the downsampled coordinate g'.

The standard coordinate decoding methods find the coordinates of the maximal and second maximal activation. The joint location is then predicted by shifting the location of the maximal activation 0.25 pixels towards the second maximal activation. This shifting is done to compensate for the quantization error. However, this is an empirical method that is found to have success but does not have any intuition behind it. Also, the predicted heatmap is not exactly Gaussian in nature and hence, the point with the maximal activation may not be estimated accurately. Hence, this work proposes a theoretically sound method to explore the entire heatmap and find the underlying maximal activation by modulating the heatmap to make it Gaussian and then using that fact along with the Taylor series to find the actual coordinates of the keypoint.

4.2.4 TransPose

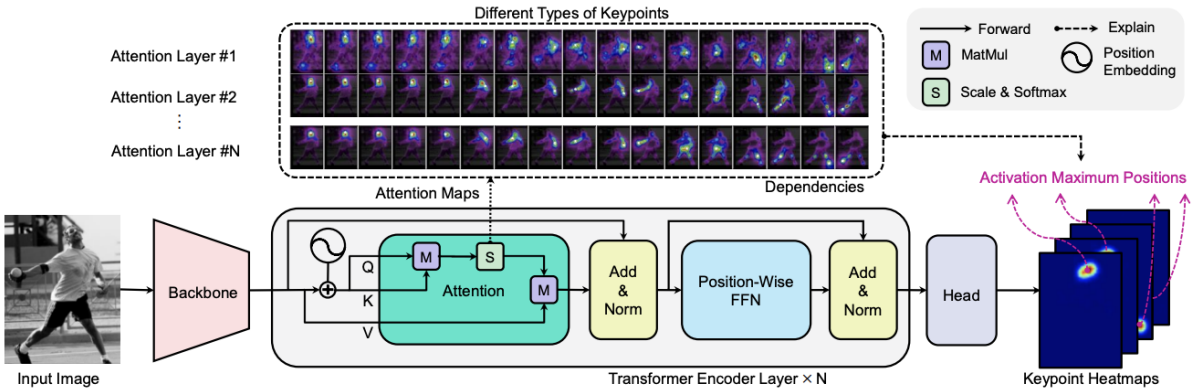


Figure 4.3: Overall Architecture of TransPose [4].

TransPose [4] is a network leverages the recent success of transformers in computer vision tasks and replicates it in pose estimation. More specifically, this network uses a common CNN backbone such as ResNet or HRNet to extract the features from the input image. The extracted features are then passed through N transformer encoder layers where each layer consists of a multi-self-attention head, layer normalization, and feed-forward neural networks. The attention layers help in capturing long-range relationships and reveal the dependencies that determine the location of the maximal activation. The N different attention layers capture different positions of maximal activation corresponding to different joints. The final attention layer acts as an aggregator, which collects contributions from image clues and forms the maximum positions of keypoints. The output from the final encoder layer is then passed through a regression head to find the heatmap for every keypoint. The overall architecture is demonstrated in Fig. 4.3.

4.3 Comparison Study

To estimate the 2D pose of the players for our custom dataset, we optimized the four SOTA 2D pose estimation networks explained in Section 4.2. We fine-tuned these SOTA models for our ice hockey dataset and compared the performance and robustness of these models.

4.3.1 Dataset

The dataset used in this chapter is the ice hockey dataset described in Chapter 3. However, we just use the human keypoints in this chapter as we are focusing on benchmarking human pose estimators.

4.3.2 Implementation Details

All of our experiments were conducted using an NVIDIA Geforce RTX 2070 GPU with 8 GB RAM. All the 2D pose estimation models were fine-tuned using the pretrained COCO [46] models for 10 epochs. Because of memory and hardware constraints, we were unable to perform extensive experiments and could only use a batch size of 8. The input size of all the models was 256×192.

To perform all our experiments, we used a 2-stage MSPN, HRNet-W32, DARK with HRnet-W32 as the backbone and TransPose with HRNet-W32 as the feature extractor. Furthermore, for MSPN, we used SGD [52] as our optimizer with a learning rate of 1e-2, momentum of 0.9, and a weight decay of 1e-5. As for all the other models, we used Adam [53] as our optimizer with a learning rate of 1e-3.

4.3.3 Results

Tables 4.1 and 4.2 provide a comprehensive overview of the training and test accuracies achieved by a network trained from the ground up. Notably, the results demonstrate that MSPN [2] outperforms other models on our dataset. A significant contributing factor to its superior performance is the effectiveness of the feature extractor. The downsampling encoder, which incorporates strategically designed residual connections along with intermediate supervision, improves feature extraction and yields high-fidelity representations. This is corroborated by our qualitative findings presented in Figs 4.4 and 4.5, where MSPN exhibits a superior ability to model the pose of a player even in the presence of occlusions, distinguishing itself from other networks.

Moreover, our analysis, as depicted in Tables 4.1 and 4.2, reveals variations in accuracy across different joints. Specifically, the shoulder, knee, and ankle joints exhibit higher accuracy compared to the wrist and elbow joints. This discrepancy underscores the challenges posed by high motion-induced blur and occlusions in the vicinity of hand joints, which are in constant, rapid motion during gameplay.

Table 4.1: Per joint and total training accuracy of the fine-tuned networks. MSPN has the highest total accuracy and highest accuracy for each joint.

Joint	**Training Accuracy**(%)			
	MSPN	**HRNet-W32**	**HRNet-W32 + DARK**	**TransPose**
Left Shoulder	97.1	95.27	95.08	91.44
Right Shoulder	96.63	95.23	95.07	91.53
Left Elbow	94.86	89.97	90.80	86.92
Right Elbow	94.91	88.33	89.52	85.53
Left Wrist	93.34	87.69	87.13	83.70
Right Wrist	92.98	86.11	84.58	81.16
Left Hip	93.76	90.52	89.07	84.78
Right Hip	94.27	90.72	88.88	84.40
Left Knee	97.19	94.38	94.70	92.59
Right Knee	97.28	94.67	94.78	92.15
Left Ankle	97.21	93.89	93.02	91.23
Right Ankle	97.40	94.18	93.92	91.40
Total accuracy	**95.71**	91.86	91.40	88.06

Table 4.2: Per joint and total test accuracy of the fine-tuned networks. MSPN has the highest total accuracy and highest accuracy for each joint.

Joint	**Test Accuracy**(%)			
	MSPN	**HRNet-W32**	**HRNet-W32 + DARK**	**TransPose**
Left Shoulder	90.13	86.13	86.92	85.83
Right Shoulder	89.96	86.23	86.98	84.95
Left Elbow	87.12	81.07	82.51	81.95
Right Elbow	89.03	81.69	85.96	84.00
Left Wrist	86.13	80.27	80.38	80.77
Right Wrist	86.61	81.18	82.26	82.04
Left Hip	79.38	75.69	75.04	74.80
Right Hip	79.75	75.63	78.45	79.50
Left Knee	92.39	89.54	89.25	88.27
Right Knee	92.76	89.35	90.66	80.81
Left Ankle	92.01	87.00	86.51	86.65
Right Ankle	93.70	87.48	86.19	86.93
Total accuracy	**88.35**	83.51	84.32	83.75



(a) HRNet (b) DARK (c) MSPN (d) TransPose-H

Figure 4.4: Inference of the implemented 2D pose estimators without occlusion.



(a) HRNet (b) DARK (c) MSPN (d) TransPose-H

Figure 4.5: Inference of the implemented 2D pose estimators with occlusion.

4.4 Conclusion

This study has effectively showcased the performance of current SOTA networks in challenging, fast-paced environments characterized by motion blur and occlusions. Through a comprehensive comparative analysis, we have delineated the strengths and weaknesses

of these existing models, revealing that hip keypoints are susceptible to errors when subjected to rapid non-linear motion. The insights gained from this research hold substantial promise for the advancement of future networks, fostering their resilience in the face of blur, occlusions, and distortions.

Chapter 5

Ensemble Model for Joint Player and Stick Pose Estimation

5.1 Introduction

In the previous chapter, we studied the advantages and disadvantages of various SOTA human pose estimators on our ice hockey dataset. Through a comprehensive analysis of the strengths and failures of such models, MSPN was demonstrated to be the best performing model for this dataset. However, these analyses showcase that hand keypoints, particularly wrists and elbows, and hip joints are susceptible to errors when subjected to rapid non-linear motion. This is mainly due to the bulky equipment and high degree of occlusions present in those joint areas.

The pose of the hockey stick held by these players can play a huge role in improving these joint accuracies. Yao *et al.* [54] and Neher *et al.* [1] showcased that modeling human-object interactions can refine the pose of the human as well. Furthermore, estimating the pose of the stick is an essential component for action recognition and player assessment. Actions that require knowledge of the stick include passing the puck and shooting the puck, as examples.

Hence, in this chapter, we develop an ensemble of two different MSPN models to jointly predict the pose of a player and their stick: one solution by training the model from scratch on all 15 keypoints (12 human + 3 stick), and the other by using a COCO pretrained MSPN network and transferring the knowledge of that network to another MSPN network that lifts these 12 human keypoints to 15 keypoints. This transfer learning paradigm is

inspired from Neher *et al.* [1]. We discuss the methodology of each individual model and the ensemble in Sections 5.2.1, 5.2.2 and 5.2.3 respectively. We then conduct extensive quantitative and qualitative studies of each model separately and the ensemble model in Section 5.3.3 to showcase the effectiveness of each component of our baseline.

5.2 Methodology

5.2.1 Training from Scratch

We use an untrained MSPN network and train it on our real-world ice hockey dataset collected from broadcast NHL videos. Since the hockey sticks are simultaneously predicted along with players, the sticks should also be captured in the cropped input image that is fed to the MSPN encoder. Hence, we leverage the *extended bounding boxes* which are calculated using the ground-truth keypoints. Finally, we estimate 15 keypoints instead of the 12 human keypoints that are generally estimated.

5.2.2 HyperMSPN: Transfer Learning

To induce additional priors that facilitate human pose estimation, we transfer the knowledge provided by large-scale human pose datasets to our model. However, standard transfer learning is not applicable as the number of keypoints in our dataset is not equal to the number of keypoints in human pose datasets. Hence, inspired from Neher *et al.* [1], we leverage 2 MSPN networks: a human keypoint feature extractor and a lifting module. The human keypoint feature extractor is a two-stage MSPN network that is pretrained on the COCO [46] human pose estimation dataset. This module extracts features related to the 12 human keypoints, and since this is pre-trained on COCO, it is robust to large changes in the human pose. The lifting module is another 2-stage MSPN network that lifts these 12 human keypoint features to 15 keypoint heatmaps. By doing so, we circumvent the issue of keypoint mismatch in COCO (12 human keypoints because the five head keypoints are disregarded) and our dataset (15 keypoints), and reliably transfer the knowledge gained from COCO to our dataset.

5.2.3 Ensemble Model

To mitigate the errors in the aforementioned methods, we create an ensemble model by blending the results of both the models with the use of a gating module. We specifically use

two learnable parameters α_1 and α_2 to represent the weights of each mode, and calculate the final predictions as follows:

$$y_e = \alpha_1 y_{fs} + \alpha_2 y_h, \tag{5.1}$$

where y_{fs} is the predictions from the "from scratch" model, y_h is predictions from HyperMSPN, α_1 and α_2 are the learnt weights of each model respectively.

5.3 Experiments

5.3.1 Implementation Details

All of our experiments were conducted using an NVIDIA Geforce RTX 2080 GPU with 12 GB RAM using a batch size of 8. The input size for this experiment was 256 $\times$ 192. We followed a heatmap-based approach, and used an MSE loss to minimize the distance between the groundtruth heatmaps and predicted heatmaps. Furthermore, we used SGD [52] as our optimizer with a learning rate of 1e-2, momentum of 0.9, and a weight decay of 1e-5. The total time for one training run is 20 hours.

5.3.2 Model Settings

For the from scratch model, we trained a 2-stage MSPN for 120 epochs. For HyperMSPN, we used 2-stage MSPN networks for both the human keypoint feature extractor and the lifting module respectively.

For the ensemble network, all the model parameters except the prediction weights α_1 and α_2 were frozen. Since the number of trainable parameters are minimal, training was conducted for 10 epochs. Furthermore, for MSPN, similar to our benchmarking experiment in Chapter 4, we used SGD [52] as our optimizer with a learning rate of 1e-2, momentum of 0.9, and a weight decay of 1e-5.

5.3.3 Results

Training From Scratch. The training and test accuracies for each joint and overall accuracy of our MSPN model trained from scratch is reported in Table 5.1. In comparison to simply predicting the human joints as shown in Table 4.2, jointly predicting the stick joints improves the test accuracy of the left and right hip joints by a significant margin of 12.56%

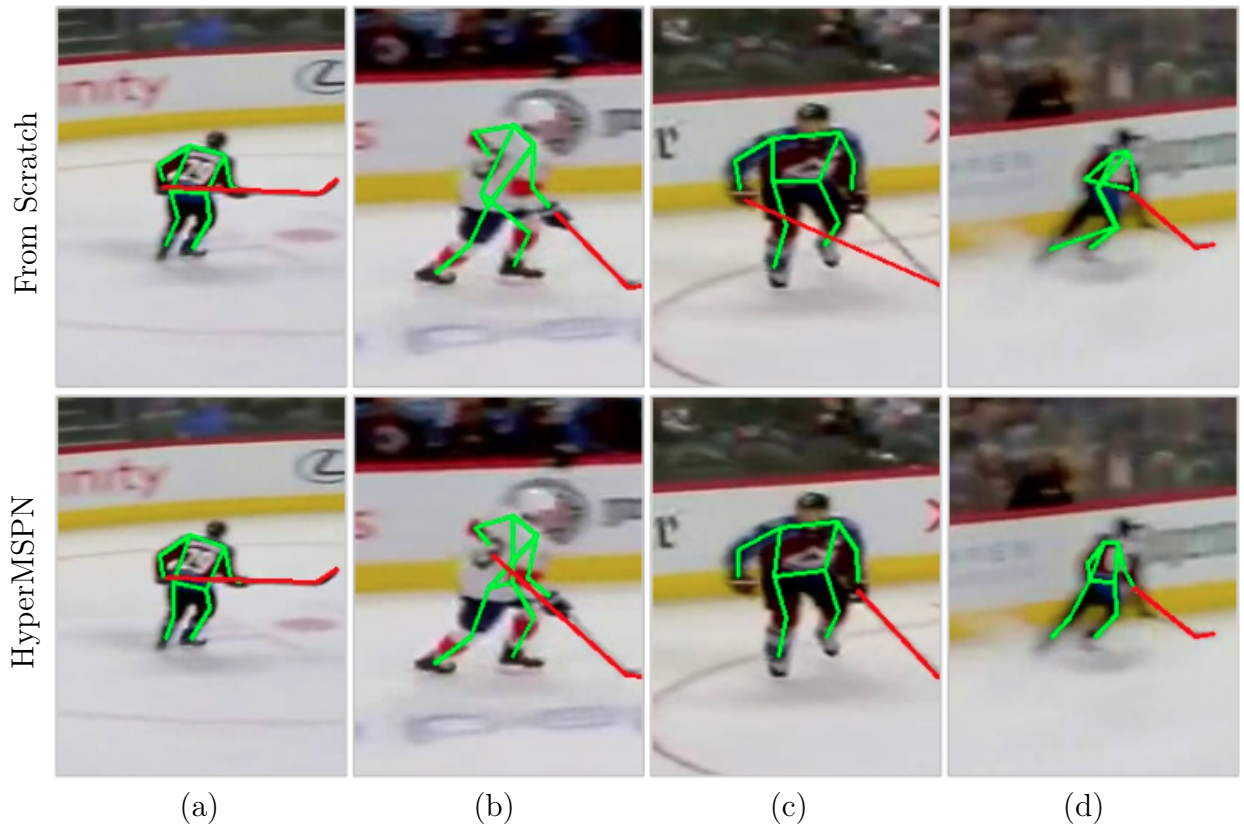


Figure 5.1: Qualitative comparison of the two MSPN models trained on our ice hockey dataset.

(from 79.38% to 91.94) and 12.11% (from 79.75% to 91.86%) respectively, thereby improving the overall accuracy of the human pose and showcasing the impact of the estimation of stick keypoints.

Furthermore, we also showcase the qualitative results of this model in Fig. 5.1. This model produces accurate results, but can yield suboptimal results in cases of high motion blur as depicted in Fig. 5.1.

HyperMSPN. We assess the performance of our proposed HyperMSPN by reporting the training and test accuracies. The results, reported in Table 5.1, demonstrate the superior performance of this method in comparison to the *from scratch* training method as the former method outperforms the latter by 2.3% on the test set. The qualitiative comparison presented in Fig. 5.1 depict the superior performance of this method as this overcomes the

Table 5.1: Quantitative comparison of our from scratch, HyperMSPN and Ensemble models. Here, FS denotes the from scratch model, H-MSPN denotes the HyperMSPN model and EM denotes the ensemble model.

Joint	**Train Accuracy**			**Test Accuracy**		
	FS	**H-MSPN**	**EM**	**FS**	**H-MSPN**	**EM**
Left Shoulder	99.32	99.21	**99.31**	91.28	93.94	**94.16**
Right Shoulder	98.99	99.02	**99.20**	91.07	93.27	**93.62**
Left Elbow	99.01	98.80	**99.18**	86.19	90.33	**91.10**
Right Elbow	98.90	98.68	**98.91**	87.04	92.44	**92.44**
Left Wrist	98.42	98.22	**98.55**	86.23	89.56	**90.98**
Right Wrist	97.79	97.81	**98.22**	86.21	**90.97**	90.59
Left Hip	99.11	99.03	**99.18**	91.94	92.55	**93.93**
Right Hip	99.17	99.09	**99.20**	91.86	93.60	**94.55**
Left Knee	99.09	99.10	**99.29**	92.94	94.79	**95.52**
Right Knee	99.01	98.96	**99.13**	92.56	94.41	**94.80**
Left Ankle	98.48	98.56	**98.82**	92.88	94.07	**94.86**
Right Ankle	98.16	98.41	**98.65**	92.87	93.56	**93.86**
Stick Butt End	**99.03**	98.92	97.21	70.30	71.73	**71.94**
Stick Heel	96.83	97.24	**97.20**	59.21	60.28	**61.59**
Stick Toe	96.23	**97.34**	96.90	55.61	57.25	**57.27**
Overall Accuracy	98.45	98.42	**98.69**	84.55	86.85	**87.41**
Overall Body Accuracy	98.72	98.56	**99.08**	90.26	92.79	**93.36**

limitations of the model trained form scratch.

Error Analysis. While the HyperMSPN quantitatively outperforms the MSPN trained from scratch and addresses the challenges faced by the latter, a deeper analysis reveals some shortcomings of the HyperMSPN model itself. Surprisingly, further investigation indicates that some errors made by HyperMSPN are complementary to those of the MSPN trained from scratch. This is illustrated in Fig. 5.2. In the first example, the from-scratch model estimated a semantically incorrect stick pose, while HyperMSPN predicted the correct pose. In the second example, the from-scratch model outperformed HyperMSPN, as the butt end of the stick and the occluded right wrist joints were not accurately captured by HyperMSPN. These findings validate our claim that combining both the models using an

ensembling technique eliminates the errors presented by each model individually.



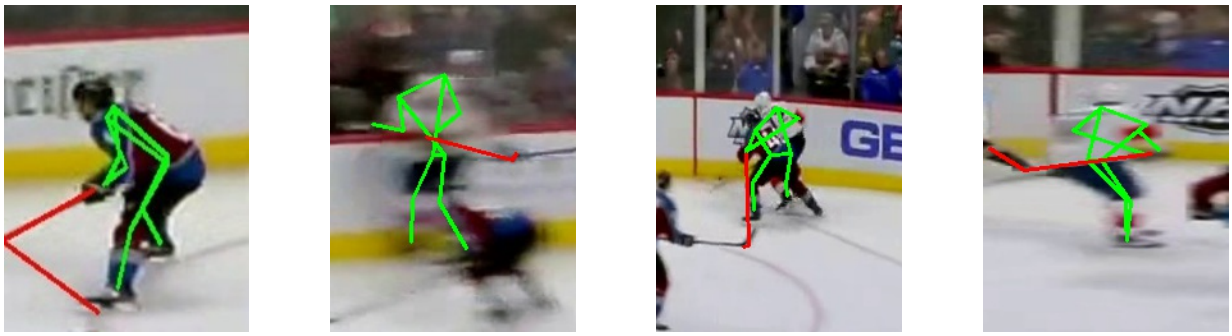
Figure 5.2: Qualitative depiction of the complimentary nature of the errors presented by from scratch model and HyperMSPN.

Ensemble Learning. The training and test accuracies of our ensemble model are highlighted in Table 5.1. From the table, it is evident that the ensemble model takes the best of both models and outperforms both the from scratch model and HyperMSPN on each

and every joint. This results in an overall performance boost of 0.56% over HyperMSPN, thereby showcasing the efficacy of our proposed ensemble model.

5.4 Limitations

While the ensemble solution presented in this section improve the quality and precision of our pose predictions significantly, they are still limited by some constraints: All the joints have to present within the image, the image has to be of reasonable quality (which is very difficult to satisfy considering the high degree of motion blur) and these methods work well only when there is one person in the input image. The aforementioned challenges are illustrated in Fig. 5.3. We solve these issues to our best extent in the next chapter.



(a) (b) (c) (d)

Figure 5.3: Limitations of our ensemble baseline solution. In (a), the stick is not entirely captures within the frame, causing the model to generate spurious predictions. In (b), the player and their hockey stick is highly distorted due to motion blur. In (c) (d), multiple players come into the image, thereby confusing the model.

Chapter 6

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

In the previous chapter, we leveraged various techniques such as ensemble learning and transfer learning to develop robust pose estimators using existing SOTA networks. However, as mentioned in Section 5.4, these networks are bound by a set of constraints which are not satisfied in most of the scenarios for ice hockey. Hence, we circumvent these constraints by creating a novel multimodal solution for out-of-bounding box keypoint prediction. In this chapter, we generally refer to the object that a human holds as their *extension*. Hence, in this chapter, *extension* and hockey stick are interchangably used. Similarly, the players are also more generally referred to as humans in this chapter.

This chapter is organized as follows: The problem statement is clearly explained and investigated in the introduction, while also providing a brief description of the solution in Section 6.1; The related works for this work are then mentioned in Section 6.2; A detailed overview of the methodology is presented in Section 6.3; The quantitative and qualitative results on three different datasets along with the ablations conducted on out ice hockey dataset are highlighted in Section 6.4; and lastly, the chapter is concluded in Section 6.5.

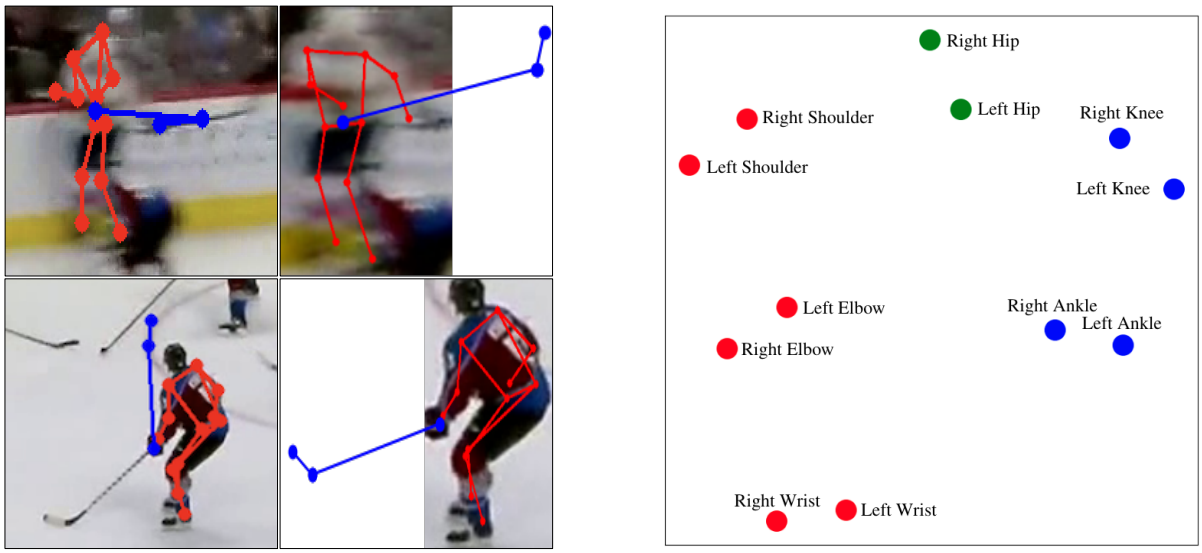
6.1 Introduction

Contemporary deep learning-based pose estimation methods predominantly follow a top-down approach: cropping each person in an image using bounding boxes before estimating their pose individually[2, 18, 21, 3, 36, 7, 38, 20]. Despite the progress in human and animal pose estimation [55, 56, 18, 38, 41], predicting the pose of humans along with their *extensions* (objects that they hold) is a problem that has not been addressed extensively in the literature. Though the ulterior objective is to extract 2D keypoints from an image, directly adapting existing pose estimators to jointly predict human pose and their extensions yields suboptimal results as shown in Fig. 6.1(a). Random keypoints such as markings on the ice rink are mispredicted as stick keypoints. Also, self-occlusions and player-player occlusions make it quite difficult to predict the stick keypoints. *While SOTA 2D pose estimators work really well on animals and humans without any major architectural changes, why do they perform poorly for human extensions?* We address and answer this question in our work.

To begin, we revisit the pipelines of SOTA 2D pose estimators. All the existing SOTA 2D pose estimators follow the top-down approach, i.e, they first detect entities of interest, crop them and then estimate their pose on the cropped images. Hence, existing networks work well on single person/object pose estimation. However, when we try to include the extensions of humans within the image, this increases the field-of-view, which in-turn introduces noise (unnecessary visual features) that was previously not present. Fig. 6.1(a) depicts the aforementioned limitation, where various unnecessary features like other players' hockey sticks, faceoff circles and people in the audience enter the field-of-view. We argue that *this increased field-of-view poses a challenge by implicitly including more noise in the image*, making it more difficult to extract robust features.

How do we reduce the field-of-view while still correctly estimating the extensions? We provide a simple yet impactful solution to fix the problem: we just do not capture the extensions. We reformulate the top-down pose estimation problem and leverage the fact that the structure of the human and their extension are highly correlated [1, 57] to predict out-of-bounding box keypoints. By doing so, we reduce the unwanted noise introduced by the widened field-of-view while capturing more robust human and extension keypoints. However, this means that we have no information about the extensions in our input image. This begs the question: *How do we provide spatial context of the missing keypoints?*

To answer this, we turn to other ways of informing the model about the *missing* keypoints. Recent works have shown that using language to induce semantic context of keypoints can lead to effective feature representations [58, 49, 59, 60]. Specifically, existing



(a) (b)

Figure 6.1: Difference between existing networks and our network. (a) Qualitative comparisons between our baseline (left) and (right) `TokenCLIPose`. (b) Visualization of keypoint-specific prompt embeddings.

methods introduced by Hu *et al.* [49] on human pose estimation align the image features with keypoint-specific text embeddings generated from vision language models (VLMs) using a contrastive loss. However, these text embeddings primarily capture local details, neglecting the crucial global relationship between lower-body and upper-body joints; In Fig. 6.1(b), language models encode similar joints together, while losing the global structure. Hence, *while text serves as a great supervisory signal, biasing the image feature extractor to the text embeddings could lead to loss of global context.*

To overcome the above challenges, we present a novel and effective solution `TokenCLIPose` to significantly improve reliability under dynamic real-world scenarios by leveraging language to *"see" beyond the bounding box.* Specifically, by combining the rich semantic representations of keypoint-specific text embeddings with image features, and a transformer to capture global dependencies, we extract superior fine-grained representations which significantly boosts the performance across the board. First, we use a backbone to extract multi-scale visual features and coarse human keypoint locations from the cropped input image. Using the rich semantic meaning provided by keypoint-specific text embeddings in conjunction with these features, a transformer decoder is then employed to capture the

global dependencies between different visual, location, and text features instead of directly aligning the visual features to the text embeddings. Furthermore, registers [61] are used in our transformer decoder to extract fine-grained spatial representations of keypoints and reduce artifacts in the feature maps.

6.2 Related Work

Occlusion-Aware Pose Estimation: Various methods tackle the problem of occlusions by studying the relationships between multiple humans present in an image [50, 48, 62]. However, we do not compare with their works as we are interested in top-down approaches. Park *et al.* [63] addresses the problem of *out-of-bounding box* keypoints by refining the bounding box before pose prediction, but this is computationally expensive. Our approach, on the other hand, provides a parameter-free technique to induce spatial context for the *out-of-bounding box* keypoints by leveraging the knowledge of VLMs.

Vision-Language Models: Language supervision has been shown to improve feature representations in various vision tasks, such as image classification, semantic segmentation, and pose estimation. Radford *et al.* [64] proposed the contrastive pretraining paradigm CLIP that leverages contrastive learning to optimize a text and image encoder jointly to classify images. This work also showcased the significance of large-scale vision-language pretraining by demonstrating accurate zero-shot classification. Following CLIP, other research [65, 66, 67] focused on improving image classification using CLIP. Zhang *et al.* [68] transfer the 2D pre-trained knowledge to 3D domains, thereby improving zero-shot point-cloud recognition. Rao *et al.* [69] showcased the efficacy of CLIP pretraining on dense prediction tasks by converting the image-text matching problem to a pixel-text matching problem.

Recently, Tevet *et al.* [60] leverage text to inpaint missing poses in a sequence in a spatiotemporally consistent manner. Guo *et al.* [59] use pose-aware prompts to predict 3D hand meshes from images. Recent methods on 2D pose estimation [49, 58] utilize joint-specific keypoints to learn richer representations. Particularly, they use a contrastive loss to align the image features extracted from a backbone to their keypoint-specific text embeddings. This, however, is suboptimal as the text embeddings do not capture global structure between human keypoints. Hence, aligning the image features to text embeddings might lead to loss of spatial information and inaccurate localization of joints. We improve this by using a transformer-based network to capture the global spatial dependencies between image and text features, thereby guiding our image features using text supervision rather than biasing them to the text prompts themselves.

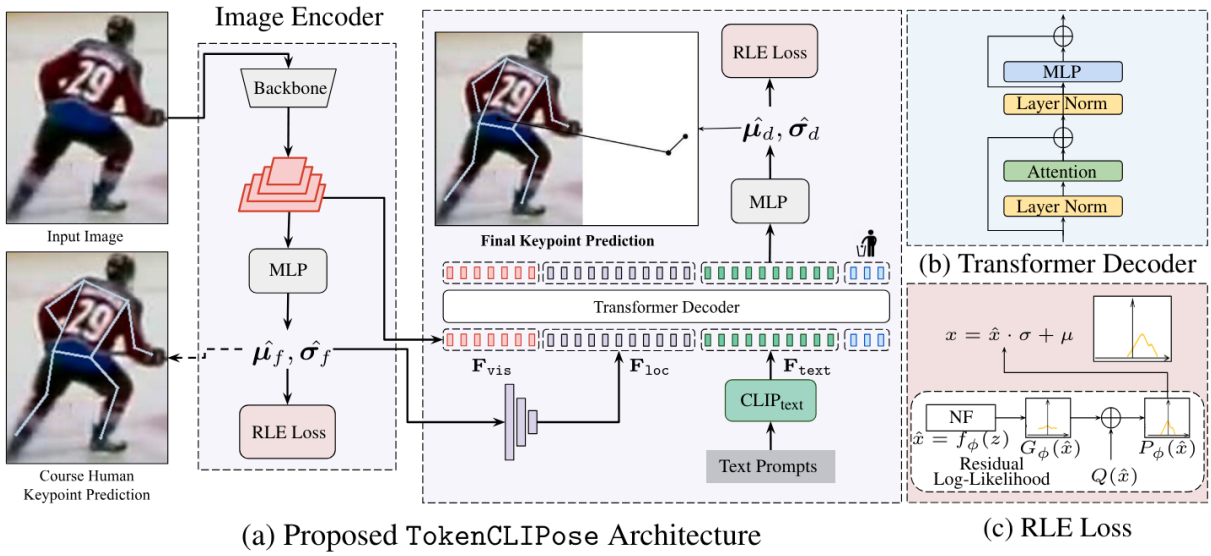


Figure 6.2: `TokenCLIPose` Architecture: We first incorporate an image encoder to extract multi-scale image features and coarse human keypoint locations $\hat{\boldsymbol{\mu}}_f$, and project them onto a joint multimodal embedding space obtaining image tokens $\mathbf{F}_{\texttt{vis}}$ and location tokens $\mathbf{F}_{\texttt{loc}}$ respectively. Then, we leverage a text-based keypoint encoder to extract keypoint-specific text tokens $\mathbf{F}_{\texttt{text}}$ from VLMs. These multimodal tokens are fed to a transformer decoder to capture spatial dependencies between them and predict all the 2D keypoints $\hat{\boldsymbol{\mu}}_d$. The coarse human keypoint predictions and the final keypoint predictions are supervised using the RLE loss.

6.3 Method

6.3.1 Problem Formulation

Given a cropped image $\mathbf{I} \in \mathbb{R}^{h \times w}$ of a human, generated from bounding boxes obtained through a detection network, we predict K_{out} keypoints $\hat{\boldsymbol{\mu}}_d \in \mathbb{R}^{K_{out} \times 2}$ that represent the 2D poses of extensions and/or humans, along with the scale parameter $\hat{\boldsymbol{\sigma}}_d$ for each keypoint.

6.3.2 Network Architecture

We propose an encoder-decoder architecture to estimate 2D keypoints that are *not captured* in the bounding box. An image encoder extracts multi-scale image features from the

cropped image, which are then passed through a multi-layer perceptron (MLP) to generate coarse human keypoint proposals. These image features and locations are then projected onto the multimodal embedding space to form image and location tokens. The text-based keypoint encoder leverages text prompts to generate keypoint-specific text tokens. Finally, all these multimodal tokens are concatenated and passed through a transformer decoder to capture global relationships between these tokens and predict final 2D keypoints. The network is illustrated in Fig. 6.2.

Image Encoder. The cropped input image is initially passed through a pretrained CNN to extract multi-scale dense feature maps. These multi-scale feature maps are fused and projected onto a joint multimodal embedding space to form the image tokens $\mathbf{F}_{\texttt{vis}} \in \mathbb{R}^{N \times C_{emb}}$, where N is the number of tokens and C_{emb} is the joint multimodal space dimension. Furthermore, they are processed through an MLP to generate coarse human keypoint predictions $\hat{\boldsymbol{\mu}}_f \in \mathbb{R}^{K_h \times 2}$, and a scale parameter $\hat{\boldsymbol{\sigma}}_f \in \mathbb{R}^{K_h \times 1}$, which are optimized using the RLE process as detailed in Section 6.3.3.

Text-based Keypoint Encoder. Following image feature extraction, the keypoint encoder tackles the challenge of *unseen/missing* keypoints in the image. We employ language-guided keypoint representations, where text prompts encode class-related information of each keypoint. By leveraging CLIP's pretrained text encoder, we generate the text tokens $\mathbf{F}_{\texttt{text}} \in \mathbb{R}^{K_{out} \times C_{emb}}$ which are then concatenated with the extracted image tokens. This process indirectly injects visual context of the missing keypoints into the model, even when it is not directly visible in the cropped input image. Additionally, we incorporate the coarse human keypoint predictions $\hat{\boldsymbol{\mu}}_f$ and convert them to location tokens by projecting them onto the joint multimodal space $\mathbf{F}_{\texttt{loc}} \in \mathbb{R}^{K_h \times C_{emb}}$, and concatenate them with the image and text tokens. The final set of tokens fed to the decoder is denoted as $\mathcal{F} = \{\mathbf{F}_{\texttt{vis}}, \mathbf{F}_{\texttt{text}}, \mathbf{F}_{\texttt{loc}}\}$.

Transformer Decoder. To predict all the keypoints precisely, the relationships between these multimodal tokens ($\mathcal{F}$) need to be captured accurately. Therefore, we leverage the transformer decoder to understand the inherent correlations between different text, image and location tokens. Instead of treating each modality separately and utilizing a cross-attention mechanism to understand the associations between them, we treat all tokens together as a homogeneous entity and employ the standard self-attention mechanism. This approach offers a simpler way to gain a holistic view of the relationships between all keypoints, locations, and image patches. The transformer layers are followed by MLPs to estimate the final keypoint predictions $\hat{\boldsymbol{\mu}}_d \in \mathbb{R}^{K_{out} \times 2}$ and the scale parameter $\hat{\boldsymbol{\sigma}}_d \in \mathbb{R}^{K_{out} \times 1}$. To reduce artifacts in feature maps and understand richer relationships, we employ additional register tokens following Darcet *et al.* [61].

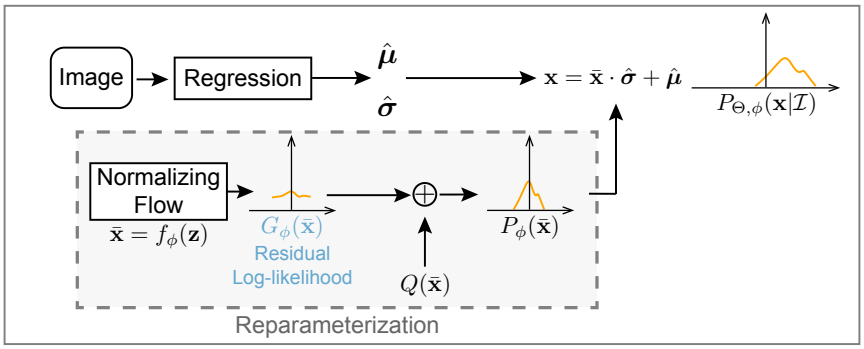


Figure 6.3: Detailed overview of Residual Log-Likelihood Loss

6.3.3 Loss Function

Following RLE and Poseur [41, 20], we formulate the regression task as a distribution learning problem and adopt maximum likelihood estimation (MLE) to effectively predict the output coordinates. We use normalizing flows to estimate the deviation in predicted and groundtruth keypoint distributions. In mathematical terms, given an input image $\mathcal{I}$, our network estimates a distribution $P_{\Theta,\Phi}(\mathbf{x}|\mathcal{I})$ representing the probability of the groundtruth keypoint appearing at the location $\mathbf{x}$. Here, Θ and Φ denote the parameters of this pose network and the flow model f_ϕ, respectively. The flow model f_ϕ acts as a refinement step, iteratively transforming a preset Gaussian distribution $\mathbf{z} \sim \mathcal{N}(0, \mathrm{Id})$ to capture the deviation of the predicted and groundtruth distributions using the network's prediction ($\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\sigma}}$).

Equation (6.1) depicts the mathematical formulation of the RLE loss, where $Q(\bar{\boldsymbol{\mu}}_g)$ is the preset Gaussian distribution, $G_\phi(\bar{\boldsymbol{\mu}}_g)$ is the learned distribution by the flow model, $\bar{\boldsymbol{\mu}}_g = (\boldsymbol{\mu}_g - \hat{\boldsymbol{\mu}})/\hat{\boldsymbol{\sigma}}$ represents the normalized difference between ground-truth and predicted keypoints, and s is a constant term. Refer Li *et al.* [41] for more details.

$$\begin{aligned} \mathcal{L}_{RLE} &= -\log P_{\Theta,\phi}(\mathbf{x}|\mathcal{I})\Big|_{\mathbf{x}=\boldsymbol{\mu}_g} \\ &= -\log P_\phi(\bar{\boldsymbol{\mu}}_g) + \log \hat{\boldsymbol{\sigma}} \\ &= -\log Q(\bar{\boldsymbol{\mu}}_g) - \log G_\phi(\bar{\boldsymbol{\mu}}_g) - \log s + \log \hat{\boldsymbol{\sigma}}. \end{aligned} \tag{6.1}$$

Similar to Mao *et al.* [20], we supervise both the coarse predictions $(\hat{\boldsymbol{\mu}}_f, \hat{\boldsymbol{\sigma}}_f)$ and the final predictions from the decoder $(\hat{\boldsymbol{\mu}}_d, \hat{\boldsymbol{\sigma}}_d)$ using RLE. Hence, our final loss function is

$$\mathcal{L} = \mathcal{L}^f_{RLE} + \mathcal{L}^d_{RLE} \tag{6.2}$$

where,

$$\mathcal{L}^{f}_{RLE} = -\log P_{\Theta_f,\Phi_f}(\mathbf{x}|\mathcal{I})\Big|_{\mathbf{x}=\boldsymbol{\mu}_g} \quad \text{and} \quad \mathcal{L}^{d}_{RLE} = -\log P_{\Theta_d,\Phi_d}(\mathbf{x}|\mathcal{I})\Big|_{\mathbf{x}=\boldsymbol{\mu}_g}$$

Here, Θ_f and Φ_f denote the parameters of the backbone network and the flow model of the coarse keypoint predictions, and Θ_d and Φ_d denote the parameters of our decoder regression model and flow model of the final keypoint predictions.

6.4 Experiments

A critical challenge in evaluating pose networks for human extensions is that there are no publicly available datasets containing both the human and extension pose annotations. To address this limitation, we create, from "scratch", two new, original sports datasets, ice hockey and lacrosse, featuring pose annotations for humans and their corresponding extensions. Experiments on these datasets demonstrate the performance of our model in estimating the pose of extension keypoints, which are not captured within the bounding box. We further evaluate the model's ability to predict unseen keypoints (due to occlusion/self-occlusion) within the bounding box using the benchmarked multi-person cluttered dataset, Crowdpose [70], validating the generalizability of our model.

For human pose estimation, the total keypoints (K_{out}) depend on extensions. Without extensions, K_{out} equals the number of human keypoints (K_h). When extensions are present, K_{out} increases to K_h plus the number of extension keypoints (K_e).

6.4.1 Ice Hockey Dataset

Training. All networks are trained for 200 epochs with a batch size of 64 on a single NVIDIA GeForce RTX 4090 GPU. We employ the Adam optimizer with a learning rate of 6×10^{-4} for all CNN-based architectures while the transformer-based architectures are trained using the AdamW optimizer with an initial learning rate of 3×10^{-4}. The weight decay is set to 10^{-5} for all models. A stepLR scheduler was used to linearly reduce the learning rate from the initial value of 10^{-5}. Consistent with [41], we utilize RealNVP [71] as the flow model within our model.

Results. The results, presented in Table 6.1, emphasize our method's *SOTA performance* over existing methods in robustly predicting the ice hockey player and the hockey stick

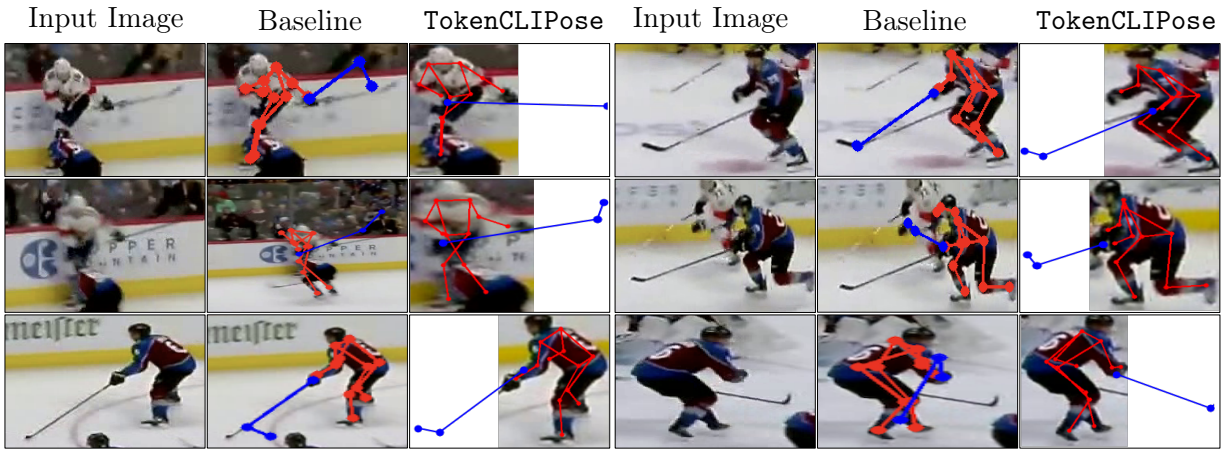


Figure 6.4: Qualitative comparison of `TokenCLIPose` with the ensemble baseline solution discussed in Chapter 5 on our ice hockey dataset.

keypoints, outperforming existing works by 2.60%. This is further validated by the qualitative comparisons in Fig. 6.4, where our model demonstrates the ability to predict semantic poses even in challenging scenarios with extreme motion blur and occlusion. These results support our hypothesis that *re-considering the extension pose estimation task as an unseen keypoint problem without explicitly capturing it in the bounding box* minimizes the impact of background information leading to accurate poses.

6.4.2 Lacrosse Dataset

Zero-shot Results. We evaluate `TokenCLIPose`'s generalization capability by performing zero-shot transfer from our pretrained model on ice hockey to the lacrosse dataset. Due to the difference in the shape of the head of a lacrosse and hockey stick, we predict two keypoints corresponding to the shaft of a lacrosse stick. Furthermore, the text prompts for the two extension keypoints are also changed while keeping the model frozen. The results presented in Table 6.2 showcase that our `TokenCLIPose` model outperforms the established baselines by 1.87%, thereby demonstrating the efficacy of our proposed model for generalizable pose estimation.

Table 6.1: **Comparison with SOTA Methods** on our real-world ice hockey dataset (PCKh@0.5). * represents the ensemble baseline that we discussed in Chapter 5. **BoldFace** represents the highest score. Underline represents the second-highest score. All models were trained with an input resolution of 256 × 192.

Method	Backbone	Body	Butt End	Heel	Toe	Overall Acc
SimpleBaseline [36]	ResNet-50	93.19	69.57	57.19	52.76	86.52
HRNet [3]	HRNet-W32	93.24	70.48	59.29	55.36	86.93
TransPose [4]	HRNet-W32	93.30	72.63	61.18	57.41	87.38
MSPN* [2]	-	93.36	71.94	61.59	57.27	87.41
ViTPose [18]	ViT-B	95.61	71.94	61.33	58.80	89.29
`TokenCLIPose`	ResNet-50	95.81	74.86	65.79	65.08	90.36
`TokenCLIPose`	HRNet-W32	97.17	75.41	66.70	66.34	91.63
`TokenCLIPose`	MSPN	**97.37**	**75.94**	**67.82**	**66.15**	**91.89**
Improvement	-	**1.76% ↑**	**4.00% ↑**	**6.49% ↑**	**7.35% ↑**	**2.6% ↑**

Table 6.2: **Zero-shot Comparison with SOTA Methods** on our real-world Lacrosse dataset (PCKh@0.5). * represents the ensemble baseline that we discussed in Chapter 5. **BoldFace** represents the highest score. Underline represents the second-highest score. All models were trained with an input resolution of 256 × 192.

Method	Backbone	Body	Butt End	Heel	Overall Acc
SimpleBaseline [36]	ResNet-50	94.73	67.28	53.99	89.85
HRNet [3]	HRNet-W32	95.84	70.68	57.40	91.29
MSPN* [2]	-	95.92	71.35	58.41	91.48
ViTPose [18]	ViT-B	95.77	72.85	60.18	91.59
`TokenCLIPose`	HRNet-W32	**97.24**	**76.60**	**65.01**	**93.46**
Improvement	-	**1.47% ↑**	**3.75% ↑**	**4.83% ↑**	**1.87% ↑**

6.4.3 CrowdPose Dataset

Training. We employ the AdamW optimizer with an initial learning rate of 6×10^{-4} and weight decay of 0.1. Following ViTPose [18], we apply linear warmup for the first 2000 iterations with a warmup factor of 10^{-3}. Furthermore, we perform gradient clipping to prevent overfitting.

Results. Table 6.3 presents a comparison between established pose estimation techniques and our method. As shown in the table, `TokenCLIPose` outperforms all the top-down approaches by 3.8% demonstrating its efficacy in predicting unseen keypoints robustly. We also depict the qualitative results of `TokenCLIPose` to depict its robustness to occlusion. It is notable that even in scenarios where we only see the side-view of humans, `TokenCLIPose` estimates the pose reliably.



Figure 6.5: Qualitative Results of our `TokenCLIPose` on the CrowdPose dataset.

Table 6.3: **Comparison with SOTA methods** on CrowdPose dataset. **BoldFace** represents the best score. Underline represents the second-best score. Different methods have different input resolutions as these aere the standard results reported directly from the respective publications.

Method	Input Resolution	AP	AP_{50}	AP_{75}	AP_E	AP_M	AP_H
Mask-RCNN [30]	800 × 800	57.2	83.5	60.3	69.4	57.9	45.8
AlphaPose [72]	-	61.0	81.3	66.0	71.2	61.4	51.1
SimpleBaseline [36]	384 × 288	60.8	81.4	65.7	71.4	61.2	51.2
CrowdPose [70]	320 × 256	66.0	84.2	71.5	75.5	66.3	57.4
Hourglass-104 [73]	384 × 288	65.2	85.9	69.5	-	-	-
KAPAO-L [34]	384 × 288	68.9	89.4	75.6	76.6	69.9	59.5
HRNet-W48 [3]	384 × 288	69.3	89.7	75.6	77.7	70.6	57.8
Transpose-H [4]	384 × 288	71.8	91.5	77.8	79.5	72.9	62.2
HRFormer-B [22]	384 × 288	72.4	91.5	77.9	80.0	73.5	62.4
`TokenCLIPose`	384 × 288	**76.2**	**93.9**	**82.4**	**83.3**	**77.4**	**66.1**
Improvement	-	**3.8% ↑**	**2.4% ↑**	**4.5% ↑**	**3.3% ↑**	**3.9% ↑**	**3.7% ↑**

6.4.4 Ablation Studies

We conduct comprehensive ablations to study the effect of our design choices and verify the impact of each module on our proposed `TokenCLIPose`. For consistency, all the ablations are conducted on our ice hockey dataset unless specified otherwise.

Gains from Each Modality. The influence of each modality on the proposed model's performance is shown in Table 6.4. As evidenced by a performance improvement of 3.35%, the inclusion of text tokens plays a significant role in enhancing the pose estimation accuracy. In contrast, we find that including the location tokens do not improve the performance significantly, only by a small margin of 0.45%.

Table 6.4: **Effect of Different Modalities** on the overall accuracy.

Text tokens ($\mathbf{F}_{\texttt{text}}$)	Location tokens ($\mathbf{F}_{\texttt{loc}}$)	Image tokens ($\mathbf{F}_{\texttt{vis}}$)	Overall Acc
✗	✗	✓	88.09
✗	✓	✓	88.54
✓	✓	✓	**91.89**

Do we need to treat each modality heterogeneously? We probe whether multimodal

tokens need to be treated heterogeneously by testing out different attention mechanisms for our transformer decoder. As shown in Table 6.5, utilizing self-attention directly on all the tokens produces similar results to performing self-attention separately on image, location and text tokens, and then applying cross-attention. Therefore, treating the tokens from each modality separately is not necessary. We hypothesize that this could be due to all the tokens being projected to the same joint embedding space, thereby eliminating the need to process them differently.

Table 6.5: **Effect of Attention Mechanisms** on the overall accuracy.

Attention-Mechanism	Overall Acc
Intention [74]	91.51
Self-attention	91.89
Cross-attention	**91.92** ↑ 0.03

Impact of Text Prompts. We investigate the influence of language on our model's efficacy on the ice hockey dataset by varying the degree and complexity of text prompts that we use. Starting from randomly initializing the text tokens instead of using CLIP embeddings, we study the effects of using single prompts and the prompt ensemble technique proposed in Radford *et al.* [64] for ImageNet classification. It is evident from Table 6.6 that incorporating text instead of randomly initializing text tokens results in a significant improvement of 2.31% in the overall accuracy. Furthermore, using prompt ensemble techniques improve the accuracy of the model by 1.04%, showcasing the importance of the quality of text prompts.

Table 6.6: **Effect of text prompts** on the stick accuracy.

Prompt type	Stick Accuracy	Overall Acc
No text	63.37	88.54
Single Prompt	67.41	90.85
Prompt Ensemble	**69.97** ↑ 2.56%	**91.89** ↑ 1.04%

Influence of Bounding Boxes. We evaluate the influence of bounding box predictions on our model's performance by using groundtruth bounding boxes obtained from groundtruth human poses and the bounding boxes from the FasterRCNN object detector. The results are showcased in Table 6.7. The table illustrates our model's robustness to the quality of

bounding boxes, as using groundtruth bounding boxes improves the model's performance by a small margin of 2.47%.

Table 6.7: **Impact of bounding boxes** on the overall accuracy.

Bounding boxes	Overall Acc
Faster-RCNN [75]	91.89
Groundtruth	**94.36** ↑ 2.47%

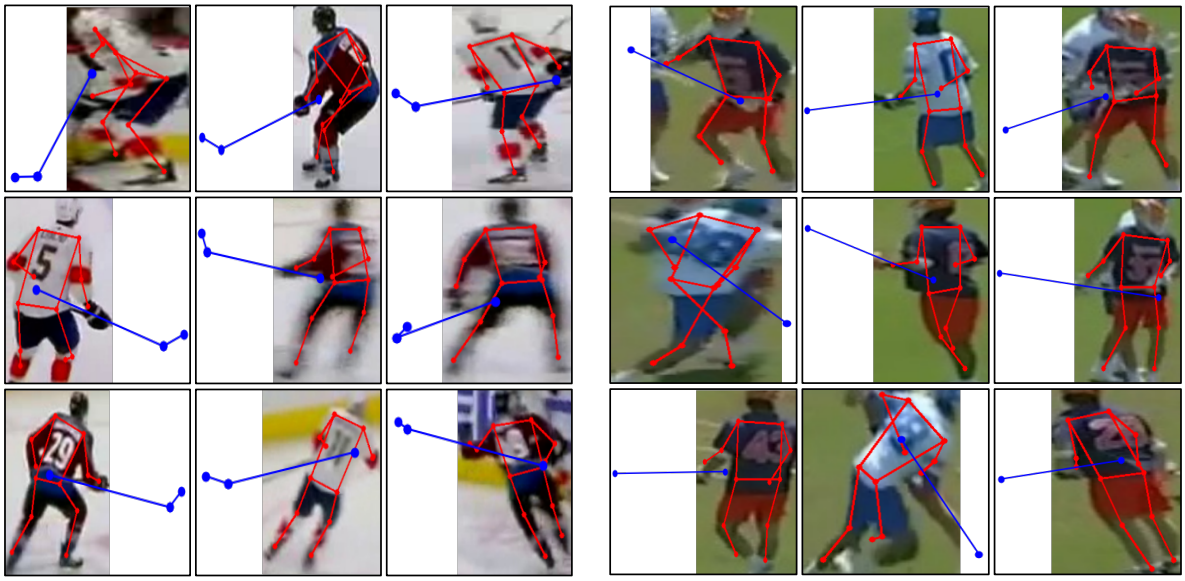
Comparison with Multimodal Pose Estimators. To investigate the effectiveness of our technique of leveraging text with existing multimodal methods, we compare `TokenCLIPose`'s performance with the multimodal counterpart LAMP [49]. Though LAMP does not follow the top-down approach, it is the only network that uses language for 2D human pose estimation. Thus, we compare and highlight the results in Table 6.8. As shown in the table, it is evident that `TokenCLIPose` outperforms LAMP on the CrowdPose dataset by 4.8%. This corroborates our claim that *leveraging text as supervisory signals to guide image features provides better performance than aligning image features to the text tokens.*

Table 6.8: **Comparison with LAMP** on CrowdPose dataset

Method	Input Resolution	AP	AP_{50}	AP_{75}	AP_E	AP_M	AP_H
LAMP [49]	512 × 512	71.4	90.3	77.1	77.9	72.1	64.2
`TokenCLIPose`	384 × 288	**76.2**	**93.9**	**82.4**	**83.3**	**77.4**	**66.1**

Additional Qualitative Results. To further illustrate the effectiveness of our approach, we present additional qualitative results on the ice hockey dataset in Fig. 6.6. Furthermore, we demonstrate the model's generalization capabilities by achieving strong performance on a Lacrosse dataset, highlighting the impact of language on the model's robustness to domain changes.

Player Handedness. After predicting the pose of the players and their hockey sticks, we determined the handedness of each player (left-handed or right-handed) in our test video through a heuristic-based post-processing step. Specifically, we estimated handedness in each frame by analyzing the y-coordinates of the wrist joints: if the left wrist had a higher y-coordinate than the right wrist, the player was classified as right-handed, and vice-versa. By applying this method across all frames, we then employed majority voting on each



(a) Ice-hockey (b) Lacrosse

Figure 6.6: Qualitative Results of `TokenCLIPose` for the extension pose estimation task on our ice hockey and lacrosse datasets. We plot the pose of extensions outside the cropped image to understand how `TokenCLIPose` works.

player tracklet (frames containing that partiular player) to determine the handedness of each unique player in the test video. This offline heuristic-based approach achieved 100% accuracy on the test set, demonstrating the effectiveness of our pose model, `TokenCLIPose`, in accurately capturing player poses.

6.5 Conclusion

In this work, we proposed `TokenCLIPose`, an innovative solution for robustly predicting human and *extension* poses. Instead of explicitly modeling extensions within the bounding box, we reformulated the *extension* pose estimation as an unseen keypoint prediction problem. We leverage the power of large pre-trained VLMs in augmenting the spatial information of unseen keypoints. We showcased that capturing relationships between multimodal tokens is more effective than aligning image features to text tokens. To evaluate the effectiveness of `TokenCLIPose` in predicting *extension* keypoints, we created real-world datasets for ice hockey and lacrosse. We significantly outperform existing top-down methods on these

datasets (by 2.60% and 1.87%, respectively). Additionally, we achieve a 3.8% improvement over prior top-down networks on the CrowdPose dataset. This shows `TokenCLIPose`'s flexibility to predict unseen keypoints within the bounding box as well.

Chapter 7

Conclusion and Future Work

2D pose estimation has been widely addressed in computer vision. Pose estimation, in general, refers to predicting the pose of an entity (humans, in most cases), and helps to determine their physical orientation with respect to an environment. This plays a vital role in sports analytics, in estimating the 'correct pose' of players to decode their style of play, analyze their actions, avoid injuries, plot their activities, and for a gamut of other use cases. In turn, it helps coaches and scouts to make better informed decisions during hiring and improve player and team performance. Vision-based pose estimation is also relevant because of their minimally invasive approach and methodology.

This thesis presents various improvements made for automated 2D pose estimation from hockey broadcast videos.

7.1 Summary of Contributions

The contributions of this thesis can be summarized as follows:

- **Benchmarking Existing Works.** In this thesis, we first introduce a new benchmark for 2D pose estimation on ice hockey broadcast videos. Leveraging the latest SOTA netowks on human pose estimation, we create a comprehensive evaluation of the best works and analyze their strengths and weaknesses to great depths and find out that MSPN produces the best results for our use-case, outperforming the other models by a significant margin.

- **MSPN-based ensemble solution for Joint Player and Stick Pose Estimation.** In order to jointly predict the pose of the player along with their stick, we train a MSPN model under three different settings: we train a MSPN model from scratch to predict the 15 keypoints; we create HyperMSPN using a unique transfer learning paradigm using 2 MSPNs where we initialize the first MSPN (human keypoint feature extractor) with COCO pre-trained weights and let the model predict just the 12 human keypoint features. The output of this first MSPN is then fed to another lifting module which lifts these 12 to 15 keypoints, thereby making use of the knowledge gained from the COCO dataset. this improves the validation accuracy by 2.3% over the from scratch model; Finally, we study the complimentary nature of the errors presented by both our trained models and hence, develop an ensemble model where we just learn the weights for both these models while keeping those models frozen, leading to an improvement of 0.56% over HyperMSPN.

- `TokenCLIPose`**: Multimodal Out-of-bounding box 2D Pose Estimation.** In order to overcome the challenges of existing top-down pose estimators, we present a simple yet effective solution to significantly improve reliability under dynamic real-world scenarios by leveraging language to *see beyond the bounding box*. Specifically, by combining the rich semantic representations of keypoint-specific text embeddings with image features, and a transformer to capture global dependencies, we extract superior fine-grained representations which significantly boosts the performance of jointly predicting the pose of the stick and the player across the board. `TokenCLIPose` outperforms our baseline by 4.48% on the ice hockey dataset.

- **New Dataset for 2D Pose Estimation in ice hockey.** To the best of our knowledge, there is only one open-source dataset for ice hockey pose estimation that is available [1]. However, that dataset consists of basic player movements in isolated conditions. However, we introduce the first pose dataset captured from NHL broadcast videos, thereby making it much more realistic and hence difficult. Additionally, we also introduce a small-scale real-world lacrosse dataset to showcase the generalization capabilities of `TokenCLIPose`.

7.2 Future Research

The work presented in this thesis provides a basis for further research into joint pose estimation of hockey player and stick from broadcast videos. Here are some of the directions that can be pursued:

7.2.1 Action recognition using joint pose

Human action recognition is one of the most important tasks for analyzing human motion. The most successful approaches model this problem as a function of pose estimation, based on the keypoint representation of human joints. However, most of them use just the pose of the human to recognize actions. However, in fields where the human interacts with an object, the pose of the object plays a vital role in understanding the pose of the human. Hence, using the pose of the human and object for player action and event recognition is expected to be immensely useful.

7.2.2 Unique player identification using player handedness

Unique player identification is another fundamental module in vision-driven sports analytics that lends its use in player assessment. The most common approach to this task is jersey number recognition. Existing methods simply use image features of a player tracklet to recognize the jersey number, thereby trying to uniquely identify the player. However, using the handedness of the player as an additional feature would aid in more accurate player identification as it reduces the search space significantly. We have showcased that we can robustly detect the handedness of a player from their pose. Hence, concatenating these handedness features with the image features could lead to robust player identities.

7.2.3 Finding Tape Color

Players in ice hockey apply tape to their blades to get a better grip of the puck so that they do not lose control when shooting, passing or stick-handling [1]. Finding the color of the stick tape can again be beneficial for unique player identification as it helps us reduce the search space further. We will need to use the pose of the hockey stick to reliably find the color of the tape on the blade. However, that is quite difficult since the blade is often occluded or blurred in the original image.

7.2.4 Temporal 2D pose estimation

Our research focuses on predicting the pose of players and hockey sticks from single RGB images. However, since we are dealing with broadcast videos, taking advantage of the

[1] https://howieshockeytape.com/blogs/howies-hockey-takes/why-do-you-tape-a-hockey-stick

temporal domain and performing pose estimation over time would help in refining the pose in frames where the player or their stick is not completely visible. This would help maintain temporal consistency which can lead to more accurate results. However, since 2D pose is directly estimated from high-dimensional data such as images, performing temporal pose estimation is highly compute intensive unlike 3D pose which is extracted from a low-dimensional input 2D pose.

7.3 Thesis Applicability

This thesis introduces a novel pose estimation network `TokenCLIPose` to predict out-of-image 2D keypoints. This can be leveraged in various domains where humans closely interact with objects, such as shovelling, tennis and soccer. Furthermore, as shown in our experiments on the CrowdPose dataset, `TokenCLIPose` also reliably predicts the pose of a human when there is high human-human interaction. Hence, this can be used for pose estimation in highly crowded scenarios, which in-turn facilitates action recognition. Hence, this can also be used for surveillance and monitoring people in crowds.

7.4 Thesis Impact

This dissertation provides a comprehensive evaluation of existing 2D pose estimators for ice hockey and a new way to reconsider the pose estimation task to predict out-of-bounding box keypoints. This thesis's contribution of `TokenCLIPose` improves our understanding of why joint hockey stick pose prediction along with player pose is difficult for existing top-down networks and presents one way to mitigate the issues. This understanding should enable the application of top-down solutions for multi-instance pose estimation which encounters similar issues. Furthermore, accurate stick pose prediction implies that teams can make more data-driven decisions to improve their teams' performance without incurring additional overhead expenses. This also implies that the use of invasive technology such as infrared sensors can be avoided to a large extent to robustly analyze players and teams.

References

[1] Helmut Neher, Kanav Vats, Alexander Wong, and David A Clausi. Hyperstacknet: A hyper stacked hourglass deep convolutional neural network architecture for joint player and stick pose estimation in hockey. *2018 15th Conference on Computer and Robot Vision (CRV)*, pages 313–320, 2018.

[2] Wenbo Li, Zhicheng Wang, Binyi Yin, Qixiang Peng, Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, and Jian Sun. Rethinking on multi-stage networks for human pose estimation. *ArXiv*, abs/1901.00148, 2019.

[3] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5686–5696, 2019.

[4] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Keypoint localization via transformer. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11782–11792, 2020.

[5] Helmut Neher, Mehrnaz Fani, David A Clausi, Alex Wong, and John Zelek. Pose estimation of players in hockey videos using convolutional neural networks. In *2017 Ottawa Hockey Analytics Conference (OTTHAC), Ottawa, Canada*, 2017.

[6] Marjan Shahi, David Clausi, and Alexander Wong. Goalienet: A multi-stage network for joint goalie, equipment, and net pose estimation in ice hockey. *ArXiv preprint arXiv:2306.15853*, 2023.

[7] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, 2016.

[8] Martin A. Fischler and Robert A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, C-22:67–92, 1973.

[9] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61:55–79, 2004.

[10] Remi Ronfard, Cordelia Schmid, and Bill Triggs. Learning to parse pictures of people. In *Computer Vision—ECCV 2002: 7th European Conference on Computer Vision Copenhagen, Denmark, May 28–31, 2002 Proceedings, Part IV 7*, pages 700–714. Springer, 2002.

[11] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[12] Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. Poselet conditioned pictorial structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–595, 2013.

[13] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, 2005.

[14] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. *CoRR*, abs/1312.4659, 2013.

[15] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. *Advances in neural information processing systems*, 27, 2014.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *ArXiv*, abs/1505.04597, 2015.

[18] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *ArXiv*, abs/2204.12484, 2022.

[19] Yuanhao Cai, Zhicheng Wang, Zhengxiong Luo, Binyi Yin, Angang Du, Haoqian Wang, Xinyu Zhou, Erjin Zhou, Xiangyu Zhang, and Jian Sun. Learning delicate local representations for multi-person pose estimation. *CoRR*, abs/2003.04030, 2020.

[20] Wei Mao, Yongtao Ge, Chunhua Shen, Zhi Tian, Xinlong Wang, Zhibin Wang, and Anton van den Hengel. Poseur: Direct human pose regression with transformers. In *European Conference on Computer Vision*, 2022.

[21] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shutao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11293–11302, 2021.

[22] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution transformer for dense prediction. *ArXiv*, abs/2110.09408, 2021.

[23] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V. Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. *CoRR*, abs/1511.06645, 2015.

[24] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*, 2016.

[25] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[26] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[27] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. *Advances in Neural Information Processing Systems*, 30, 2017.

[28] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5386–5395, 2020.

[29] Guillem Brasó, Nikita Kister, and Laura Leal-Taixé. The center of attention: Center-keypoint grouping via attention for multi-person pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11853–11863, 2021.

[30] Kaiming He, Georgia Gkioxari, and Piotr Dollár. Girshick ross. mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.

[31] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: A simple and strong anchor-free object detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):1922–1933, 2020.

[32] Zhi Tian, Hao Chen, and Chunhua Shen. Directpose: Direct end-to-end multi-person pose estimation. *ArXiv preprint arXiv:1911.07451*, 2019.

[33] Weian Mao, Zhi Tian, Xinlong Wang, and Chunhua Shen. Fcpose: Fully convolutional multi-person pose estimation with dynamic instance-aware convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9034–9043, 2021.

[34] William McNally, Kanav Vats, Alexander Wong, and John McPhee. Rethinking keypoint representations: Modeling keypoints and poses as objects for multi-person human pose estimation. *ArXiv preprint arXiv:2111.08557*, 2021.

[35] Jie Yang, Ailing Zeng, Shilong Liu, Feng Li, Ruimao Zhang, and Lei Zhang. Explicit box detection unifies end-to-end multi-person pose estimation. In *International Conference on Learning Representations*, 2023.

[36] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. *ArXiv*, abs/1804.06208, 2018.

[37] Junjie Huang, Zheng Zhu, Feng Guo, and Guan Huang. The devil is in the details: Delving into unbiased data processing for human pose estimation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5699–5708, 2019.

[38] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7091–7100, 2019.

[39] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv preprint arXiv:2010.11929*, 2020.

[40] Xiao Sun, Bin Xiao, Shuang Liang, and Yichen Wei. Integral human pose regression. *ArXiv*, abs/1711.08229, 2017.

[41] Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. Human pose regression with residual log-likelihood estimation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11005–11014, 2021.

[42] Ke Li, Shijie Wang, Xiang Zhang, Yifan Xu, Weijian Xu, and Zhuowen Tu. Pose recognition with cascade transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1944–1953, June 2021.

[43] Paschalis Panteleris and Antonis Argyros. Pe-former: Pose estimation transformer. In *International Conference on Pattern Recognition and Artificial Intelligence*, pages 3–14. Springer, 2022.

[44] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *ArXiv*, abs/2005.12872, 2020.

[45] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *ArXiv*, abs/2010.04159, 2020.

[46] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'a r, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[47] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, Yizhou Wang, and Yonggang Wang. Ai challenger : A large-scale dataset for going deeper in image understanding. *ArXiv*, abs/1711.06475, 2017.

[48] Mu Zhou, Lucas Stoffl, Mackenzie Weygandt Mathis, and Alexander Mathis. Rethinking pose estimation in crowds: Overcoming the detection information bottleneck and ambiguity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14689–14699, October 2023.

[49] Shengnan Hu, Ce Zheng, Zixiang Zhou, Chen Chen, and Gita Reese Sukthankar. Lamp: Leveraging language prompts for multi-person pose estimation. *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3759–3766, 2023.

[50] Rawal Khirodkar, Visesh Chari, Amit Agrawal, and Ambrish Tyagi. Multi-instance pose networks: Rethinking top-down pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3122–3131, 2021.

[51] Tao Jiang, Xinchen Xie, and Yining Li. Rtmw: Real-time multi-person 2d and 3d whole-body pose estimation. *ArXiv preprint arXiv:2407.08634*, 2024.

[52] Sebastian Ruder. An overview of gradient descent optimization algorithms. *CoRR*, abs/1609.04747, 2016.

[53] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

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

[55] Chen Li and Gim Hee Lee. Scarcenet: Animal pose estimation with scarce annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17174–17183, 2023.

[56] Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. Ap-10k: A benchmark for animal pose estimation in the wild. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

[57] Georgia Gkioxari, Ross B. Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8359–8367, 2017.

[58] Xu Zhang, Wen Wang, Zhe Chen, Yufei Xu, Jing Zhang, and Dacheng Tao. Clamp: Prompt-based contrastive learning for connecting language and animal pose. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2023.

[59] Shaoxiang Guo, Qing Cai, Lin Qi, and Junyu Dong. Clip-hand3d: Exploiting 3d hand pose estimation via context-aware prompting. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4896–4907, 2023.

[60] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023.

[61] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *ArXiv preprint arXiv:2309.16588*, 2023.

[62] Dongkai Wang and Shiliang Zhang. Contextual instance decoupling for robust multi-person pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11060–11068, June 2022.

[63] Soonchan Park and Jinah Park. Localizing human keypoints beyond the bounding box. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1602–1611, 2021.

[64] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[65] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024.

[66] Yixuan Wei, Yue Cao, Zheng Zhang, Houwen Peng, Zhuliang Yao, Zhenda Xie, Han Hu, and Baining Guo. iclip: Bridging image classification and contrastive language-image pre-training for visual recognition. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2776–2786, 2023.

[67] Renrui Zhang, Zhang Wei, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Jiao Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. *ArXiv*, abs/2207.09519, 2022.

[68] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8552–8562, 2022.

[69] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022.

[70] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10863–10872, 2019.

[71] Laurent Dinh, Jascha Narain Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *ArXiv*, abs/1605.08803, 2016.

[72] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *IEEE/CVF International Conference on Computer Vision*, 2017.

[73] Jia Li, Wen Su, and Zengfu Wang. Simple pose: Rethinking and improving a bottom-up approach for multi-person pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11354–11361, 2020.

[74] Marta Garnelo and Wojciech Marian Czarnecki. Exploring the space of key-value-query models with intention. *ArXiv preprint arXiv:2305.10203*, 2023.

[75] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015.