Shaan Chanchani

HIST30701

Prof Mendon-Plasek

May 10, 2025

1. The Points That Define Us (and Animals)

Keypoint estimation, the task of identifying the coordinates of predefined points in images or video, has become a cornerstone technology within Artificial Intelligence. From the face filters on social media that precisely track our eyes and mouths, to advanced robotics systems navigating complex environments, to scientific tools analyzing animal behavior through pose estimation, the ability to locate consistent, meaningful points is ubiquitous.

The keypoint schemes widely used today are deeply rooted in historical decisions, technical limitations, and practical compromises made by researchers over decades. Consider the Common Objects in Context (COCO) dataset's 17-point human pose model—now a standard in the field.[1] These specific points weren't selected arbitrarily but emerged from a complex interplay of what was anatomically significant, computationally feasible, and practically useful for the applications of the time. Yet when similar schemes are extended to dramatically different subjects—like the AP-10K dataset[2] that applies a modified COCO-like 17-keypoint model across 53 anatomically diverse animal species from giraffes to bison—the historical contingency of these human-centric frameworks becomes apparent. A keypoint configuration optimized for human bodies may not capture the most relevant anatomical or functional features of a quadruped or a bird. This raises fundamental questions about representation that transcend purely technical considerations.

This research proposal explores: How did we arrive at our current conventions for selecting

[1] Tsung-Yi Lin et al., "Microsoft COCO: Common Objects in Context," in *Computer Vision – ECCV 2014* (Springer, 2014).

[2] Hang Yu et al., "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).

keypoints, and what historical trade-offs were made between competing goals? These trade-offs include anatomical accuracy (placing points on actual joint centers), labeler consistency (choosing points that human annotators can reliably identify), functional relevance (selecting points that matter for downstream applications like biomechanics), and computational feasibility (limiting the number of points to what systems could process efficiently). What constitutes a "good" keypoint scheme from a historical perspective—especially when moving beyond humans?

Understanding this history reveals how AI researchers conceptualized bodies and objects over time, showing a path often driven by available data, computational limitations, and specific applications rather than by consistent, clearly articulated principles. By examining this history, we gain insight into the epistemic and subjective choices embedded in these representation schemes. This awareness is crucial for developing more thoughtful approaches to representing diverse biological forms, potentially improving comparative biomechanics research, ethological studies, and even addressing questions of fairness and representation in AI systems.

2. How Keypoint Definitions Shaped Conceptions of Progress

The history of keypoint detection begins primarily with facial analysis, where the challenge of tracking specific facial points dates back to the early days of computer vision. Long before deep learning transformed the field, researchers were developing systems to automatically locate distinctive facial features. In the 1990s, early approaches like Active Shape Models (ASMs) by Cootes et al. emerged,[3] using statistical models of shape variation to locate facial landmarks. These were followed by Active Appearance Models (AAMs),[4] which incorporated both shape and texture information. These model-based approaches required researchers to make explicit decisions about which points to track—typically choosing points around the eyes, nose, mouth, and face contour that were both anatomically significant and visually distinctive. In the early 2010s, researchers

[3] T. F. Cootes et al., "Active Shape Models - Their Training and Application," *Computer Vision and Image Understanding* 61, no. 1 (1995).

[4] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor, "Active Appearance Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, no. 6 (2001).

struggled to clearly measure advancements in facial landmark detection due to inconsistent labeling and evaluation methods.

A major turning point came with the 2013 "300 Faces In-The-Wild Challenge" (300-W), introduced by Sagonas et al.,[5] the first facial landmark detection competition using in-the-wild images. The key contribution was that this benchmark dataset unified many existing facial databases under one consistent annotation method, adopting the Multi-PIE 68-point scheme. The adoption of this specific scheme addressed the fragmentation in the field with researchers using incompatible configurations (LFPW with 35 points, AFLW with 21, HELEN with 194), making cross-database experiments nearly impossible. The Multi-PIE scheme points represented a careful balance between fiducial points corresponding to clear anatomical features (eye corners, nose tip) and ancillary points that completed the shape without precise anatomical correlates (jawline, face contour). In Sagonas et al.'s 2016 paper "300 faces In-the-wild challenge: Database and results", the authors shared results from a dedicated experiment where they had multiple expert annotators label the same set of facial images.[6] The findings were illuminating: "This experiment shows that the agreement level among the annotators is high for the landmarks that correspond to the eyes and mouth," attributing this to the fact that "these landmarks are located to facial features which are very distinctive across all human faces".[7] Conversely, for points lacking such clear visual-semantic anchors, the outcome was different: "the standard deviation is high for landmarks that do not have a clear semantic meaning. The chin is the most characteristic example of this category, as it demonstrates the highest variance".[8]

This empirical demonstration was crucial. It provided concrete evidence that the "ground truth" itself possessed varying degrees of reliability depending on the chosen keypoint. The semantic clarity of a point directly correlated with the ability of human labelers to mark it consistently. This realization had immediate practical consequences for the 300-W Challenge, as the authors

[5] Christos Sagonas et al., "300 Faces in-the-Wild Challenge: The First Facial Landmark Localization Challenge," in *2013 IEEE International Conference on Computer Vision Workshops* (2013).

[6] Christos Sagonas et al., "300 Faces In-The-Wild Challenge: Database and results," *Image and Vision Computing* 47 (2016).

[7] Sagonas et al., 9.

[8] Sagonas et al., 9.

concluded it was "more reliable to report the performance of landmark localization techniques using the 51-point mark-up (after removing the points of the face's boundary)".[9] This decision, driven by an understanding of labeler limitations, implicitly acknowledged that progress in automatic detection should perhaps be first measured on points where the target itself is unambiguously defined. Thus, the authors established that any meaningful keypoint scheme must consider not only its desired anatomical or functional relevance but also the practical realities of human annotation capabilities and the inherent distinctiveness of the features being marked.

3. Why This Matters: Parallel Evolution and Potential Cross-Pollination

While facial landmark detection was converging toward standardization through efforts like the 300-W Challenge,[10] human pose estimation experienced a similar pattern of fragmentation followed by eventual standardization. Before COCO[11] introduced its now-standard 17-point scheme in 2016, at least seven distinct public keypoint layouts were in common use by researchers, ranging from minimal 8-point upper-body models to 16-point full-body schemes. Early human pose estimation work never converged on a single skeletal representation. Datasets such as Buffy Stickmen (2008),[12] ETHZ Upper-body (2009),[13] LSP (2010),[14] FLIC (2013),[15] Penn Action (2013),[16]

[9] Sagonas et al., "300 Faces In-The-Wild Challenge: Database and results," 9.

[10] Sagonas et al., "300 Faces in-the-Wild Challenge: The First Facial Landmark Localization Challenge."

[11] Lin et al., "Microsoft COCO: Common Objects in Context."

[12] Vittorio Ferrari et al., "Groups of Adjacent Contour Segments for Object Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, no. 1 (2008).

[13] Marcin Eichner and Vittorio Ferrari, "Better Appearance Models for Pictorial Structures," in *British Machine Vision Conference* (BMVA Press, 2009).

[14] Sam Johnson and Mark Everingham, "Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation," in *British Machine Vision Conference* (BMVA Press, 2010).

[15] Ben Sapp and Ben Taskar, "MODEC: Multimodal Decomposable Models for Human Pose Estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2013).

[16] Yongmian Zhang et al., "Modeling Temporal Interactions with Interval Temporal Bayesian Networks for Complex Activity Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, no. 10 (2013).

HumanEva-I (2007),[17] and MPII Human Pose (2014)[18] each introduced their own keypoint configurations, reflecting different research priorities and practical constraints. This proliferation of incompatible schemes created significant challenges—cross-dataset training was difficult, evaluation metrics varied widely, and meaningful comparison of progress across research groups was nearly impossible. COCO's adoption of a 17-point scheme marked a pivotal moment of standardization,[19] much like the 300-W Challenge did for facial landmarks.[20] However, a critical gap in our understanding remains: to what extent did the empirical findings from facial landmark standardization[21] influence the development of human pose keypoint schemes? The 300-W Challenge explicitly demonstrated through empirical testing that keypoints with clear semantic meaning (like eye corners) had higher inter-annotator reliability than those without (like chin or face contour points). Did similar considerations inform COCO's selection of its 17 keypoints, or was its scheme primarily derived from MPII's 16-point model[22] with minimal critical re-evaluation? This research proposes a detailed examination of the citation networks and methodological decisions connecting these influential datasets.

4. Cross-Species Application: When Human Standards Meet Animal Bodies

The challenges of keypoint definition and standardization, so evident in facial and human pose estimation, become even more pronounced when these human-centric frameworks are extended to the vast diversity of the animal kingdom. The development of datasets like AP-10K[23] exempli-

[17] Leonid Sigal, Alexandru O. Balan, and Michael J. Black, "HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion," *International Journal of Computer Vision* 87, no. 1 (2010).

[18] Mykhaylo Andriluka et al., "2D Human Pose Estimation: New Benchmark and State of the Art Analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2014).

[19] Lin et al., "Microsoft COCO: Common Objects in Context."

[20] Sagonas et al., "300 Faces in-the-Wild Challenge: The First Facial Landmark Localization Challenge."

[21] Sagonas et al., "300 Faces In-The-Wild Challenge: Database and results."

[22] Andriluka et al., "2D Human Pose Estimation: New Benchmark and State of the Art Analysis."

[23] Yu et al., "AP-10K: A Benchmark for Animal Pose Estimation in the Wild."

fies this complex intersection of practical necessity and representational compromise. As Jiang et al. note, "compared to human pose estimation... animal pose estimation is still at a preliminary stage," facing "intractable challenges" such as the "huge domain shift between each species" and "scarce datasets."[24] The AP-10K dataset, introduced by Yu et al. (2021),[25] emerged as a significant effort to address these issues, becoming "the latest and largest labeled dataset for general animal pose estimation," encompassing 10,000 labeled images across 23 animal families and 54 species, each annotated with 17 landmarks.[26] Critically, the choice of these 17 keypoints for AP-10K was directly influenced by existing human pose standards, particularly the COCO dataset.[27] Jiang et al. highlight that in AP-10K,[28] "The labels are consistent with the COCO dataset labels (Lin et al., 2014b)," referring to the work by Lin et al.,[29] and that this "decision made explicitly to facilitate 'transfer learning easily between the human dataset and animal dataset (Cao et al., 2019)'", a concept explored by Cao et al..[30][31] This strategy is a pragmatic response to the "scarcity of the annotated animal pose datasets [which] hinders the use of well-developed human pose/shape estimation model structures in supervised fashion."[32] By adopting a COCO-like scheme, researchers could leverage the "prior knowledge from numerous labeled human data," as human and many quadruped anatomies share some "morphological similarity."[33] However, this pragmatic approach of transferring a human-derived keypoint scheme to such a wide array of animal anatomies raises questions reminiscent of the findings from the 300-W Challenge regarding keypoint semantics.[34] While certain COCO keypoints (e.g., "eye," "nose") might retain relatively clear semantic meaning across many mammalian species in AP-10K, the consistent and unambiguous definition of points

[24] Le Jiang et al., "Animal Umat: A Universal Animal Model for Motion Analysis and Tracking," Publication details not verified, 2024, 1.

[25] Yu et al., "AP-10K: A Benchmark for Animal Pose Estimation in the Wild."

[26] Jiang et al., "Animal Umat: A Universal Animal Model for Motion Analysis and Tracking," 8.

[27] Lin et al., "Microsoft COCO: Common Objects in Context."

[28] Yu et al., "AP-10K: A Benchmark for Animal Pose Estimation in the Wild."

[29] Lin et al., "Microsoft COCO: Common Objects in Context."

[30] Jinkun Cao et al., "Cross-Domain Adaptation for Animal Pose Estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2019).

[31] Jiang et al., "Animal Umat: A Universal Animal Model for Motion Analysis and Tracking," 8.

[32] Jiang et al., 2.

[33] Jiang et al., 13, 3.

[34] Sagonas et al., "300 Faces In-The-Wild Challenge: Database and results."

like "shoulder," "elbow," "hip," or "knee" becomes increasingly fraught when applied to the diverse morphologies of, for instance, a giraffe versus a dog, or even within different breeds of the same species. Jiang et al. echo this concern directly in their discussion of AP-10K: "However, due to the considerable differences in physical characteristics between species and between animals and humans, it still remains to be seen whether the label entirely consistent with the human dataset (Lin et al., 2014a)[35] will meet the requirements for animal research."[36] Indeed, Cao et al. (2019),[37] referenced by Jiang et al., "pointed out that the differences in defined 'bones' between each keypoint would also cause a great domain shift even among similar species, let alone between human and animals."[38] In essence, for many animal species within a broad dataset like AP-10K, a significant portion of the 17 keypoints risk becoming analogous to the "ancillary points" of the 300-W facial dataset—points whose precise anatomical correspondence or visual distinctiveness is less clear, potentially leading to lower inter-annotator agreement and a less stable "ground truth."

5. Research Proposal: Uncovering the Rationale Behind the Points

This research critically investigates the historical development of keypoint selection conventions within Artificial Intelligence. The central research problem is to uncover how specific keypoints were chosen for influential datasets, what trade-offs were made between anatomical accuracy, labeler consistency, functional relevance, and computational feasibility, and what the epistemic and practical implications of these historical choices are, particularly when these conventions are extended from humans to diverse animal forms. Understanding this history is of paramount significance because it illuminates the often unexamined assumptions and subjective choices embedded within foundational AI representations. These choices directly impact the performance and fair-

[35] Lin et al., “Microsoft COCO: Common Objects in Context.”
[36] Jiang et al., “Animal Umat: A Universal Animal Model for Motion Analysis and Tracking,” 8.
[37] Cao et al., “Cross-Domain Adaptation for Animal Pose Estimation.”
[38] Jiang et al., “Animal Umat: A Universal Animal Model for Motion Analysis and Tracking,” 3.

ness of AI systems and the validity of scientific conclusions drawn from AI-driven analyses in fields like comparative biomechanics or ethology. Neglecting this historical context risks perpetuating potentially flawed or biased representational schemes, thereby hindering scientific progress and the development of more robust and equitable AI.

This study builds upon and synthesizes existing scholarship from the computer vision field. Key works include Sagonas et al.'s (2016) empirical research on facial landmark reliability in the 300-W Challenge, which demonstrated the critical link between a keypoint's semantic clarity and inter-annotator agreement.[39] The credibility of their argument rests on rigorous experimental validation with multiple annotators. This research will explore the extent to which such empirically grounded principles influenced subsequent major standardizations in human pose (e.g., Lin et al.'s COCO dataset[40]) and animal pose estimation (e.g., Yu et al.'s AP-10K dataset,[41] as discussed by Jiang et al.[42]). These peer-reviewed publications, along with dataset documentation (e.g., from the COCO and AP-10K websites), form a core set of primary and secondary sources. The research also considers works like Cao et al. (2019), which articulate the pragmatic motivations for transfer learning that often drove the adoption of human-centric schemes for animals.[43] My work extends this by systematically examining the historical discourse and decision-making processes, questioning not just the technical transferability of models but the historical and epistemological justification for transferring specific keypoint rationales across domains.

To explore these questions, the primary methodology will be historical analysis of key documents. This includes: (1) In-depth review of the seminal peer-reviewed papers introducing influential datasets (e.g., COCO,[44] AP-10K,[45] 300-W,[46] MPII Human Pose[47]). (2) Examination of publicly available annotation guidelines, technical reports, and workshop proceedings (e.g., from

[39] Sagonas et al., "300 Faces In-The-Wild Challenge: Database and results."

[40] Lin et al., "Microsoft COCO: Common Objects in Context."

[41] Yu et al., "AP-10K: A Benchmark for Animal Pose Estimation in the Wild."

[42] Jiang et al., "Animal Umat: A Universal Animal Model for Motion Analysis and Tracking."

[43] Cao et al., "Cross-Domain Adaptation for Animal Pose Estimation."

[44] Lin et al., "Microsoft COCO: Common Objects in Context."

[45] Yu et al., "AP-10K: A Benchmark for Animal Pose Estimation in the Wild."

[46] Sagonas et al., "300 Faces in-the-Wild Challenge: The First Facial Landmark Localization Challenge"; Sagonas et al., "300 Faces In-The-Wild Challenge: Database and results."

[47] Andriluka et al., "2D Human Pose Estimation: New Benchmark and State of the Art Analysis."

CVPR, ECCV, ICCV) associated with these datasets, often found on project websites or academic archives. (3) A systematic analysis of citation networks to trace the documented influence of key findings (like those from Sagonas et al.[48]) on subsequent dataset designs and rationales. Information that may prove difficult to access, but would be valuable, includes unpublished internal discussions or early-stage design documents from the research labs involved. The "data" for this research consists primarily of textual and contextual evidence of decision-making processes. This requires specialized knowledge of computer vision history and dataset creation methodologies. While direct collaboration is not planned, this research engages with the historical self-reflection present in the work of computer vision scholars.

Potential challenges include the scarcity of explicit, detailed rationales for every historical keypoint choice, as some decisions may have been implicit or undocumented. This will be mitigated by focusing on inferring priorities from published materials, stated goals, and community discussions, while clearly acknowledging interpretive limitations. Establishing direct causal links between research threads (e.g., facial analysis influencing human pose standards) can also be difficult. The research will address this by carefully analyzing documented lines of influence (citations, shared authors, workshop themes) and distinguishing between demonstrated and inferred connections.

Tentatively, this research anticipates finding that current keypoint conventions are largely the product of path dependency, where early successes, computational limitations, data availability (often human-centric), and pragmatic needs for standardization and model transferability heavily influenced choices. These factors often led to trade-offs that prioritized annotator consistency and computational feasibility, sometimes at the expense of strict anatomical fidelity or functional relevance for diverse subjects. An additional research problem emerging from this investigation is to assess the downstream impact of these historical representational choices on current scientific research outputs in fields like comparative biology and to explore methodologies for developing more contextually appropriate keypoint schemes for non-human subjects.

[48] Sagonas et al., “300 Faces In-The-Wild Challenge: Database and results.”

6. Conclusion: Re-evaluating the Points That Define Us and Our AIs

This essay has embarked on a historical inquiry into the selection of keypoints in AI, a practice fundamental to how machines perceive and represent bodies. The central research problem—understanding how historical trade-offs in defining these "meaningful points" have shaped current AI representations, especially across the human-animal divide—is critical. These choices, far from being purely technical, are laden with epistemic assumptions that influence scientific inquiry, technological development, and questions of fairness. The consequences of not addressing this problem include the uncritical propagation of potentially biased or suboptimal representational frameworks, which can limit both scientific insight and the capabilities of AI systems.

Our historical exploration highlighted pivotal moments and enduring challenges. The work surrounding the 300-W Challenge, particularly the findings by Sagonas et al. (2016) on the importance of semantic clarity for reliable facial landmark annotation, offered a crucial lesson in the interplay between perception, definition, and data quality.[49] However, the degree to which such empirically derived insights explicitly informed the standardization of human pose estimation, epitomized by the COCO dataset,[50] remains an area ripe for deeper historical investigation. The subsequent extension of these human-centric schemes to the vast anatomical diversity of the animal kingdom, as seen in datasets like AP-10K,[51] further underscores the tension. While pragmatic motivations like facilitating transfer learning in data-scarce contexts are understandable (as noted by Jiang et al.[52] and Cao et al.[53]), this approach risks imposing ill-suited anatomical frameworks onto non-human subjects, potentially compromising the integrity of the "ground truth" for many species-specific analyses.

The historical evidence tentatively suggests that our contemporary keypoint conventions

[49] Sagonas et al., “300 Faces In-The-Wild Challenge: Database and results.”

[50] Lin et al., “Microsoft COCO: Common Objects in Context.”

[51] Yu et al., “AP-10K: A Benchmark for Animal Pose Estimation in the Wild.”

[52] Jiang et al., “Animal Umat: A Universal Animal Model for Motion Analysis and Tracking.”

[53] Cao et al., “Cross-Domain Adaptation for Animal Pose Estimation.”

are a complex legacy of path dependency, technical pragmatism, the drive for standardization, and an evolving understanding of what constitutes a "good" representation. Trade-offs were consistently made, often prioritizing annotator agreement and computational efficiency, sometimes over nuanced anatomical or functional relevance, particularly when moving beyond well-studied human forms. Returning to the essay's introduction, the "points that define us (and animals)" are indeed not immutable truths but are constructed through specific historical, technical, and socio-cultural processes. The "historical contingency" and the "epistemic and subjective choices" embedded within these schemes are crucial to acknowledge.

One might argue that these pragmatic choices were necessary for the rapid advancement of the field. While there is validity to this perspective, a historical understanding allows us to move beyond simply accepting these compromises as inevitable. It empowers us to critically evaluate their continued appropriateness and to identify where new, more thoughtful approaches are required. The significance of this historical problem, therefore, lies in its capacity to inform a more reflexive and responsible AI development paradigm—one that questions the foundations of its data and representations. For historians, this inquiry reveals how scientific communities negotiate standards and grapple with representation. For AI practitioners and other academic communities, it underscores the need for interdisciplinary dialogue to ensure that our powerful tools are built upon representations that are not only computationally tractable but also conceptually sound and ethically considered.

Further research stemming from this problem could delve into more detailed case studies of dataset creation, conduct ethnographic analyses of current annotation practices, or foster collaborations between AI researchers and domain experts (e.g., comparative anatomists, ethologists) to co-design keypoint schemes tailored to specific biological questions and diverse forms. Ultimately, understanding the history of how we choose to represent the world in our AI systems is essential for building technologies that are more accurate, equitable, and truly intelligent.

References

Andriluka, Mykhaylo, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. “2D Human Pose Estimation: New Benchmark and State of the Art Analysis.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014.

Cao, Jinkun, Hongyang Tang, Hao-Shu Fang, Xiaoyong Shen, Cewu Lu, and Yu-Wing Tai. “Cross-Domain Adaptation for Animal Pose Estimation.” In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019.

Cootes, T. F., C. J. Taylor, D. H. Cooper, and J. Graham. “Active Shape Models - Their Training and Application.” *Computer Vision and Image Understanding* 61, no. 1 (1995).

Cootes, Timothy F., Gareth J. Edwards, and Christopher J. Taylor. “Active Appearance Models.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, no. 6 (2001).

Eichner, Marcin, and Vittorio Ferrari. “Better Appearance Models for Pictorial Structures.” In *British Machine Vision Conference*. BMVA Press, 2009.

Ferrari, Vittorio, Loic Fevrier, Frédéric Jurie, and Cordelia Schmid. “Groups of Adjacent Contour Segments for Object Detection.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, no. 1 (2008).

Jiang, Le, Caleb Lee, Divyang Teotia, and Sarah Ostadabbas. “Animal Umat: A Universal Animal Model for Motion Analysis and Tracking.” Publication details not verified, 2024.

Johnson, Sam, and Mark Everingham. “Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation.” In *British Machine Vision Conference*. BMVA Press, 2010.

Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. “Microsoft COCO: Common Objects in Context.” In *Computer Vision – ECCV 2014*. Springer, 2014.

Sagonas, Christos, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. “300 Faces In-The-Wild Challenge: Database and results.” *Image and Vision Computing* 47 (2016).

Sagonas, Christos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. “300 Faces in-the-Wild Challenge: The First Facial Landmark Localization Challenge.” In *2013 IEEE International Conference on Computer Vision Workshops*. 2013.

Sapp, Ben, and Ben Taskar. “MODEC: Multimodal Decomposable Models for Human Pose Estimation.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013.

Sigal, Leonid, Alexandru O. Balan, and Michael J. Black. “HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion.” *International Journal of Computer Vision* 87, no. 1 (2010).

Yu, Hang, 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.

Zhang, Yongmian, Yifan Zhang, Eran Swears, Natalia Larios, Ziheng Wang, and Qiang Ji. “Modeling Temporal Interactions with Interval Temporal Bayesian Networks for Complex Activity Recognition.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, no. 10 (2013).