Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Composite Dataset Training for Improved Pose Detection in the Wild

Bachelor thesis
at Research Group Knowledge Technology, WTM
Dr. Philipp Allgeuer

Department of Informatics
MIN-Faculty
Universität Hamburg

submitted by
Leon Trochelmann
Course of study: Computer Science
Matrikelnr.: 7312028
on
23.05.2023

Examiners: Dr. Philipp Allgeuer
Hassan Ali

Abstract

Human Pose Detection is a fundamental computer vision task that allows computers to detect a human being in a much more sophisticated way than a simple bounding box. The field has seen great strides in recent years with models becoming increasingly sophisticated and reaching better and better scores on common benchmarks. The underlying goal is to obtain more precise and robust detection of in-the-wild data: Data reflecting real-world use cases which is free from potential biases that the popular datasets may be subject to. To minimise such biases and improve performance on in-the-wild data, we present composite dataset training: A method to reconcile differences between datasets so that a model can be trained on all of their data. The performance of such models is evaluated on various validation sets and compared to that of models trained on individual datasets. A comparative analysis is also conducted between models trained on a composite dataset and models trained on individual datasets of equal size. This analysis aims to explore the impact of dataset-external variety versus dataset-internal variety. The results show that a model trained on a composite dataset always outperforms models trained on the individual parts. Furthermore, the composite-trained models outperform models trained on individual datasets of the same size during validation on external data, demonstrating that external variety can lead to better generalisation than internal variety.

Zusammenfassung

Die Erkennung der menschlichen Pose ist eine grundlegende Aufgabe im Bereich der Computer Vision, die es Computern ermöglicht, einen Menschen auf eine viel anspruchsvollere Weise als nur mit einem einfachen Begrenzungsrahmen zu erkennen. In den letzten Jahren hat dieser Bereich große Fortschritte gemacht, wobei Modelle immer ausgefeilter werden und bessere Ergebnisse auf gängigen Benchmarks erzielen. Das übergeordnete Ziel besteht darin, eine präzisere und robustere Erkennung von Daten in realen Anwendungsszenarien zu erreichen, die frei von möglichen Verzerrungen sind, denen populäre Datensätze unterliegen könnten. Um solche Verzerrungen zu minimieren und die Leistung bei Daten in realen Szenarien zu verbessern, präsentieren wir das Training mit kombinierten Datensätzen: eine Methode, um Unterschiede zwischen Datensätzen auszugleichen, so dass ein Modell mit allen Daten trainiert werden kann. Die Leistung solcher Modelle wird anhand verschiedener Validierungssätze bewertet und mit der von Modellen verglichen, die auf den einzelnen Datensätzen trainiert wurden. Des Weiteren wird eine vergleichende Analyse zwischen Modellen, die auf einem kombinierten Datensatz und Modellen, die auf einzelnen Datensätzen gleicher Größe trainiert wurden, durchgeführt. Diese Analyse zielt darauf ab, den Einfluss von externer Vielfalt im Datensatz im Vergleich zu interner Vielfalt zu untersuchen. Die Ergebnisse zeigen, dass ein Modell, das auf einem kombinierten Datensatz trainiert wurde, immer besser abschneidet als Modelle, die auf den einzelnen Teilen trainiert wurden. Darüber hinaus übertreffen die mit kombinierten Datensätzen trainierten Modelle Modelle, die auf einzelnen Datensätzen gleicher Größe trainiert wurden, bei der Validierung mit externen Daten. Dies zeigt, dass externe Vielfalt zu einer besseren Verallgemeinerung führen kann als interne Vielfalt.

Contents

Chapter 1

Introduction

Human pose detection is a fundamental problem in computer vision that involves determining the spatial configuration of a human body in an image or video [3]. Accurate and robust human pose detection has numerous applications, ranging from activity recognition and human-computer interaction to robotics, biomechanics and animation [17][16][1]. Significant progress has been made over the years in developing pose detection algorithms, driven by advancements in deep learning and the availability of large-scale annotated datasets [19].

However, challenges persist in achieving accurate and reliable pose detection in real-world scenarios due to factors like occlusions and underrepresented poses. These challenges are exacerbated by biases present in popular datasets, which can limit the generalisation of pose detection models to external data. To address this issue, this work aims to improve performance on such data by combining datasets and mitigating the impact of biases associated with individual datasets.

In order to create such composite datasets, the annotations from different datasets that follow diverse representations of the human body must be reconciled. A unified representation to which many annotations can be converted is introduced, allowing for the seamless integration of diverse datasets. Suitable datasets for the implementation and evaluation of this approach are large-scale datasets that are already known to generalise quite well to external data, such as AI Challenger [21], COCO [13] and Crowdpose [12], which were used for this project.

The results of a range of experiments show that models trained on composite datasets consistently outperform models trained on the individual parts. They even outperform models trained on an equivalent volume of the individual datasets when evaluating on external data, demonstrating that combining data from multiple datasets allows the model to indeed generalise better than if more data from a single dataset is used.

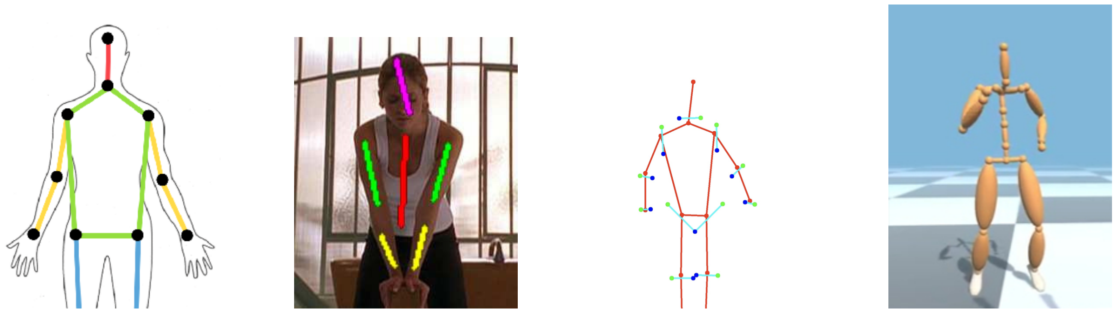
Chapter 2

Background

This section provides a concise overview of the relevant research that forms the foundation and inspiration for this work. It outlines the key research questions that will be addressed and provides an explanation of their significance within the context of this work.

2.1 Pose Detection

Pose detection, also known as pose estimation, is a task in the field of computer vision wherein the positions of specific parts of a person, animal or object are inferred from visual input. The case where the subject of the detection is a person can then be called human pose detection.
Human pose detection can be further divided based on the specific way the human body is represented. Representation can take on vastly different forms, as is illustrated by figure 2.1.



(a) AI Challenger 2D keypoint representation [21].

(b) Segmentation based stick man representation [6].

(c) MPII-TRB: Triplet representation [5].

(d) VNect 3D Human Pose estimation [15].

Figure 2.1: Differing representations in Human pose detection.

Datasets Human pose detection research heavily relies on annotated datasets for training and evaluation purposes. Several notable datasets have played a crucial role in advancing the field and assessing the performance of pose detection algorithms.
The COCO dataset is one of the most widely used datasets for human pose detection. It consists of a large collection of images annotated with human keypoints, providing a diverse range of visual scenarios and complex pose variations [13]. These qualities enable models trained on it to generalise well to new data, making it and the following datasets particularly popular.
The MPII Human Pose dataset is another prominent dataset that offers rich and diverse annotations for human pose detection [2]. Additionally, the AI Challenger (AIC) dataset [21] and Crowdpose dataset [12] are two more recent additions. AI Challenger provides a significant amount of annotated images for training and evaluation. Crowdpose focuses on pose detection in crowded scenes, offering annotations for multiple individuals within a single image.

Models Human pose detection models are usually based on convolutional neural networks [9], which come to bear in two prominent architectures: Hourglass networks [17] and the HRNet [19].
Hourglass networks, introduced by Newell et al., employ a multi-stage approach where the input image is processed through a series of down-sampling and up-sampling modules [17]. HRNet, proposed by Sun et al., focuses on maintaining high-resolution representations throughout the network [19]. These model architectures have demonstrated impressive results in human pose detection, surpassing previous approaches on challenging benchmarks such as COCO [13] and MPII [2].

Metrics In the field of human pose detection, the evaluation of models heavily relies on well-established metrics such as the Percentage of Correct Keypoints (PCK) and mean Average Precision (mAP). PCK measures the percentage of correctly detected keypoints based on a fixed distance threshold, providing a rough assessment of correctness [18].
On the other hand, mAP is a somewhat more sophisticated metric that captures the actual precision of the detections. When used for human pose detection, the metric is based on Object Keypoint Similarity (OKS) [13]. The OKS calculates the similarity between predicted keypoints and the ground truths [21].
One of the variables used to compute OKS is the sigma value: A per-keypoint factor that scales proportionally with the resulting mAP. These sigmas are particularly important for this work, as they are calculated as the standard deviation of the ground truth annotations [21], meaning they are different between datasets. This doesn't allow for a direct comparison of mAP values between datasets, as a dataset with lower sigmas generally produces a lower mAP and vice versa.

2.2 Related Work

Composite Datasets This thesis was inspired by Lambert et al. [11], who reconciled taxonomies and resolved incompatible annotations to improve generalisation capabilities for semantic image segmentation. They were able to show that a model trained on their unified semantic representation consistently performs well across domains. Models trained on their dataset achieved state-of-the-art performance on WildDash [22], a benchmark for robust semantic segmentation, without training on any of that benchmark's data, demonstrating great generalisation to new data. Similarly to their approach, incompatible annotations are reconciled for better generalisation in this work.

Kuznetsova et al. [10] have already shown the merits of composite datasets on computer vision tasks such as image classification, object detection and visual relationship detection previously. In their work, they introduced a large unified image dataset for various computer vision tasks. They were able to show that models trained on that dataset consistently outperform the respective baselines.

Despite the benefits composite datasets have shown on these tasks, they have never been applied to and evaluated on human pose detection to the best of the author's knowledge. In this work, new composite datasets are formed from existing datasets and evaluated through models trained on them to ascertain whether they can improve generalisation capabilities for pose detection.

Generalisation Capabilities Training datasets exhibit biases, and when deployed in real-world scenarios, trained models encounter data that differs from what was observed during training [20]. All datasets exhibit a degree of such a selection bias [8]. Addressing this challenge falls within the realm of generalisation, where models are expected to perform effectively in previously unseen environments after being trained on data with a great amount of variety. This new data can also be called in-the-wild data, derived from the description ”in the wild” as a setting [14].

While most popular datasets like COCO aim to include as much internal variety as possible, maintaining a small amount of bias is inevitable. This is demonstrated in this work where datasets are combined to compensate for each other's biases to allow models trained on them to achieve better performance on new data. The comparative benefits of such internal and cross-dataset external variety are a key consideration in this work.

2.3 Research Questions

The research questions for this work naturally emerge from the described state-of-the-art and adjacent research. Firstly, we seek to evaluate whether composite datasets can lead to improved performance on human pose detection challenges over the individual dataset baselines. Secondly, we perform a comparative analysis to determine whether any such improvements are merely a result of the increase in dataset volume or truly a benefit of reducing selection bias. Lastly, we analyse which aspects of the data the resulting models struggle with to identify worthwhile directions for future research.

Chapter 3

Methods

This section provides insights into the specific methods used to construct composite datasets. It outlines the datasets utilised, the training techniques applied, and the models trained on these datasets. The implementation of this project was carried out within the MMPose toolkit, a comprehensive framework for human pose detection [4].

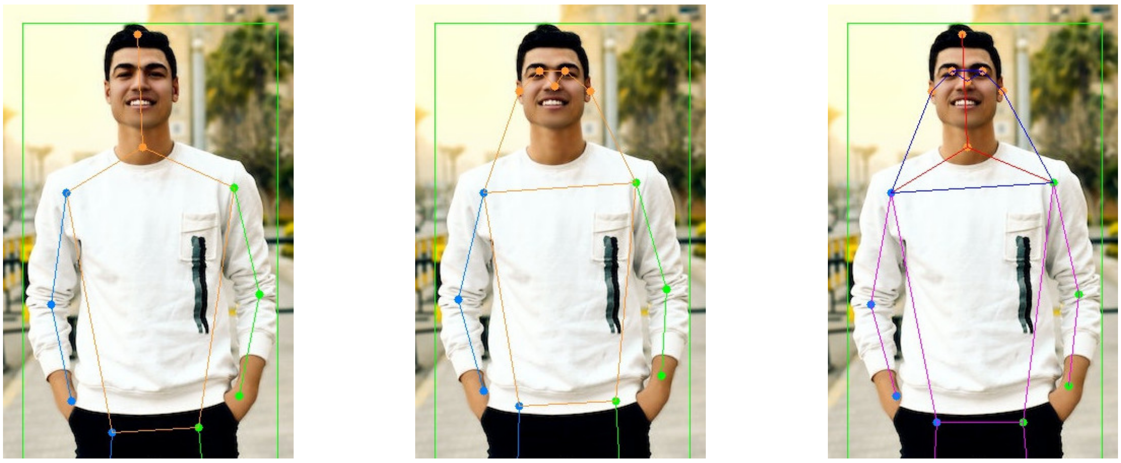
3.1 Composite Datasets

Rather than forming one very large composite dataset like Lambert et al. [11], we present a general method for forming composites for human pose detection from any arbitrary 2D keypoint detection datasets. The main challenge standing in the way of this is that the differences in the keypoint representations must be reconciled. This is achieved by making alterations to their annotations so that they follow the same style. They can then be treated as a single dataset during training.

Candidate datasets must have both ground truth person keypoint annotations and ground truth bounding box annotations to allow for building composites. The datasets COCO, AIC and Crowdpose fulfil these conditions and are opportune for this work because they share a similar annotation style. Hence they form the ideal components for the implementation of composites and their evaluation.

To treat annotations from different datasets as one and the same dataset, the annotations need to be stylistically identical. In particular, this means that the same keypoints need to be in the same order, which requires a remapping operation. This presents a problem in the case of one dataset's representation having keypoints that the other doesn't.
This can be solved by defining a unified representation which includes all keypoints from all of the individual datasets. We leverage the fact that keypoints may be flagged as "not annotated" when they are not visible or out of frame for the respective image. These keypoints are already being ignored during training and evaluation.

Hence, during the remapping operation, keypoints missing from one dataset but present in the unified representation are marked as ”not annotated” and consequently ignored. A model trained on the resulting composite dataset predicts all keypoints in the unified representation. This is illustrated by figure 3.1 for the respective representations of AIC and COCO. It's worth noting that Crowpose uses the same representation as AIC.



(a) Demonstration of AIC-style keypoint representation.

(b) Demonstration of COCO-style keypoint representation.

(c) Demonstration of the unified keypoint representation.

Figure 3.1: Illustration of the different representations side by side. The unified representation is implemented as the union of all AIC and COCO keypoints. The skeleton for the unified representation is drawn with AIC's connections in red, COCO's connections in blue, and common connections in purple.

3.2 Training

To ensure comparable results, the project utilised the same model architecture and training pipelines throughout. This allows for making conclusions on the resulting performances based solely on the selected training data. Figure 3.2 illustrates the models that were trained and the volumes of data from the different datasets used. Data was deterministically sampled from the datasets with a random seed of 0 to evaluate specific volumes and analyse their scaling with the results. A composite-trained model was always trained on the same sampled images as the models trained on its parts.

The training process itself is consistent with standard configurations included in MMPose. These configurations were only altered in certain sections to implement composite datasets and otherwise left unchanged.

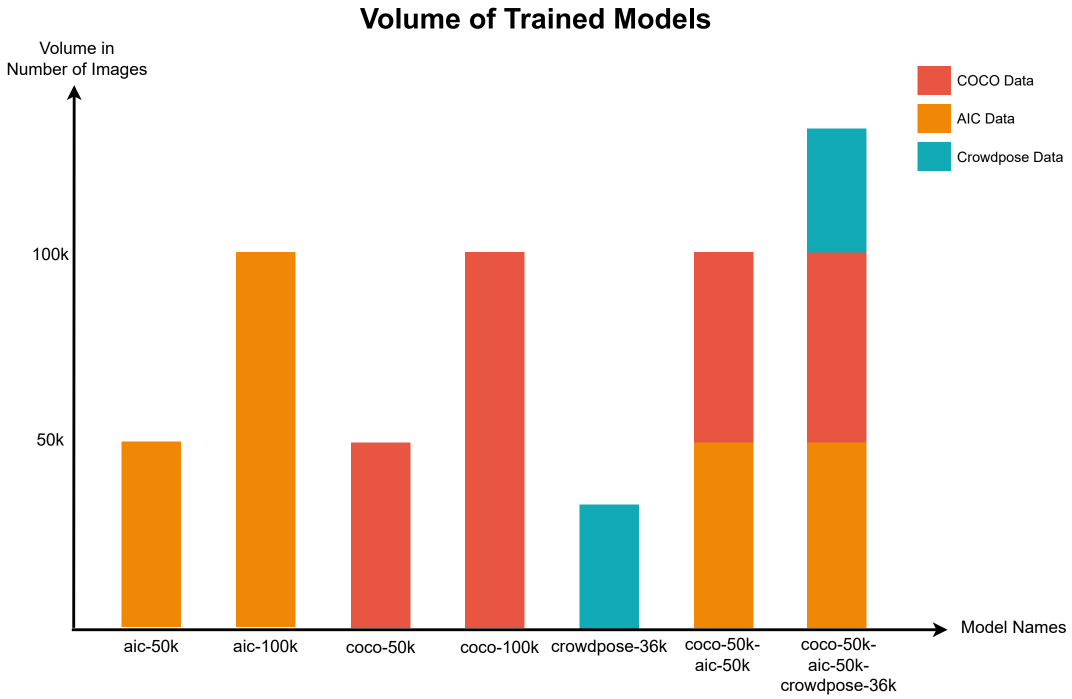


Figure 3.2: Illustration of the trained models and their volume.

Generally, the backbone architecture used for training is the HRNet with a width of 32 [19]. Each model was trained with images at a resolution of 384x288 pixels for a total of 210 epochs with a learning rate decay and a mean squared error loss.

Chapter 4

Experiments and Results

This section includes descriptions of the conducted experiments and the corresponding results as well as a discussion of these results. Each experiment is first depicted in isolation until finally an overarching discussion addressing the research questions is given at the end.
Evaluation is first conducted on labelled data from the validation sets of AIC, COCO and Crowdpose, and then unlabelled stock photography sourced from Pexels.com, representing true in-the-wild data.

4.1 Labelled Data

Fairly evaluating the varying representations and sigma values is a challenge. Since the sigmas scale with mAP, mAP between different datasets and different sigmas can not be directly compared. Hence mAP results can only be compared if they were calculated on the same keypoints and the same sigmas.
The experiments are therefore fundamentally divided by the datasets on which the models were evaluated. We also differentiate between the keypoints considered during evaluation: First, we describe evaluation on the body keypoints, which are identical in both representations and then the same is done for the head keypoints, which differ between representations.

4.1.1 Body Keypoints

The evaluation was performed with all trained models on the validation sets of AIC, COCO and Crowdpose respectively. The metric used is mAP. The results are displayed in table 4.1.

The composite models consistently outperform models trained on the parts of the respective composite. This indicates that adding more data to a model will generally improve its performance. It's worth noting that this is the case despite the datasets having different sigmas, which reflects a difference in the human ground truth annotation precision.

Evaluation on body keypoints					
Evaluation on AIC validation data		Evaluation on COCO validation data		Evaluation on Crowdpose validation data	
Model	mAP	Model	mAP	Model	mAP
coco-50k	0.1822	coco-50k	0.7342	coco-50k	0.7207
coco-100k	0.2103	coco-100k	0.7586	coco-100k	0.7629
aic-50k	0.3106	aic-50k	0.6710	aic-50k	0.7064
aic-100k	0.3355	aic-100k	0.6949	aic-100k	0.7359
crowdpose-36k	0.2693	crowdpose-36k	0.7390	crowdpose-36k	0.7300
aic-50k-coco-50k	0.3136	aic-50k-coco-50k	0.7521	aic-50k-coco-50k	0.7774
aic-50k-coco-50k-crowdpose-36k	0.3204	aic-50k-coco-50k-crowdpose-36k	0.7728	aic-50k-coco-50k-crowdpose-36k	0.7879

Table 4.1: Evaluation results when evaluating body keypoints on different validation sets.

Unsurprisingly, models trained on higher volumes of the same dataset also outperform the models trained on lower volumes.

It is also apparent that the individual datasets outperform the composites of equal volume on their own validation sets. However, the evaluation of models trained on high volumes of AIC and COCO in comparison to the model trained on a composite of AIC and COCO when evaluated on the Crowdpose validation set is particularly insightful. The composite-trained model outperforms them on this external dataset.
This indicates that a model trained on a composite dataset does indeed generalise better to external data than a model trained on a single dataset of the same volume. This in turn indicates that the composite-trained model is indeed capable of compensating for the selection bias of the individual datasets.

4.1.2 Head Keypoints

A similar evaluation to the one on the body keypoints was conducted for the head keypoints. The composites with datasets using different head representations do not add any new head data to the model, but they were evaluated regardless to monitor any possible improvements caused by internal connectivity. The results are displayed in table 4.2.

The model trained on composite data from AIC and Crowdpose outperforms models trained on the parts of the composite in this experiment as well. A composite that doesn't provide any new data doesn't appear to consistently improve performance. The minor difference in accuracy can be attributed to random factors.

Evaluation on head keypoints					
Evaluation on AIC validation data		Evaluation on Crowdpose validation data		Evaluation on COCO validation data	
Model	mAP	Model	mAP	Model	mAP
				coco-50k	0.7835
				coco-100k	0.7945
aic-50k	0.2572	aic-50k	0.8649		
aic-100k	0.2676	aic-100k	0.8794		
crowdpose-36k	0.2310	crowdpose-36k	0.9033		
aic-50k-coco-50k	0.2606	aic-50k-coco-50k	0.8991	aic-50k-coco-50k	0.7769
aic-50k-coco-50k-crowdpose-36k	0.2632	aic-50k-coco-50k-crowdpose-36k	0.9114	aic-50k-coco-50k-crowdpose-36k	0.7788

Table 4.2: Evaluation results when evaluating head keypoints on different validation sets. Models were not evaluated on keypoints that they don't predict.

4.2 In-the-Wild Data

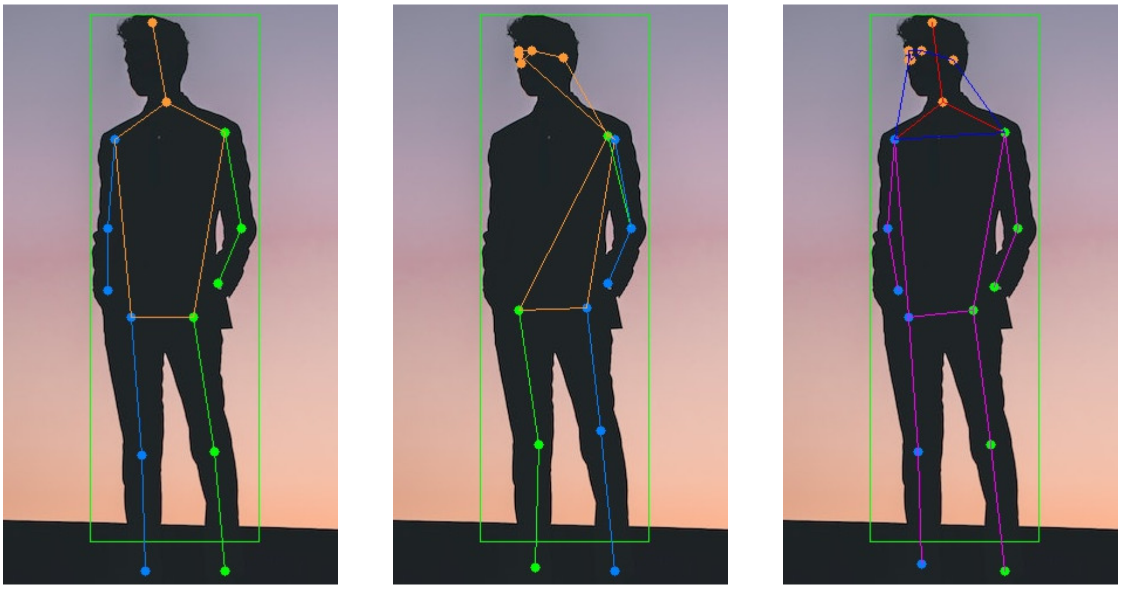
To investigate how well the models generalise on true in-the-wild data and validate the previous results, the performance of the models coco-50k, aic-50k and aic-50k-coco-50k was manually analysed on unlabelled stock photography. To obtain the required human bounding boxes, we used YOLOX as a detector [7].

50 images and 10 Videos consisting of a total of 5915 frames were passed through the models. This data was selected arbitrarily but with a focus on including a diverse array of ethnicities, body types, clothing, backgrounds and poses. The results were manually reviewed for bad detections and differences in performance between models. Since only the body keypoints were trained on data from both COCO and AIC in this case, only the body keypoints were considered during analysis.

4.2.1 Images

The results on the in-the-wild images demonstrate that none of the models seems to particularly struggle with the differences in ethnicity, body types, clothing or backgrounds. However, it appears that they do struggle with rare poses such as yoga poses, as well as partial occlusions.

Overall, the results show that the COCO-trained model is consistently outperformed by the AIC-trained model and the composite-trained model, as is exemplified in figure 4.1. This may be due to COCO's greater sigma values, reflecting lower ground truth annotation precision.



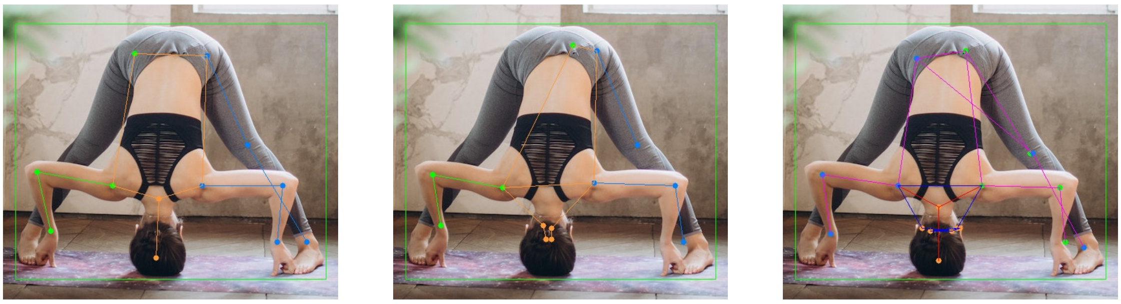
(a) Detection by the aic-50k model.

(b) Detection by the coco-50k model.

(c) Detection by the aic-50k-coco-50k model.

Figure 4.1: Pose detection with different models on unlabelled data, first image. The AIC-trained model and the composite-trained model detect the pose well, while the COCO-trained model makes a mistake.

As for the AIC-trained model and the composite-trained model, there are instances of superior detections on either side. However, there is one particularly devastating sample where the AIC-trained model confuses the sides of the person entirely, whereas the composite-trained model doesn't. This is displayed in figure 4.2.



(a) Detection by the aic-50k model.

(b) Detection by the coco-50k model.

(c) Detection by the aic-50k-coco-50k model.

Figure 4.2: Pose detection with different models on unlabelled data, second image. Keypoints of the left side of the body are drawn in green and those of the right side in blue. The composite-trained model identifies the left and right sides of the body correctly for the most part, while the individually trained models confuse the sides entirely.

All in all these results hint at a slight improvement of the composite-trained model over the individually trained ones, but show no clear winner due to the low sample size. They do however serve a purpose in demonstrating that rare and difficult poses are the primary remaining challenge for pose detection models trained on these datasets.

4.2.2 Videos

The videos paint a much clearer picture of the comparative performance in the wild between models. They contain many frames of difficult poses during activities such as dancing. These poses push the models to their limits and show a small but clear improvement of the composite-trained model over the others. The initial frames of a breakdancing video are a prime example. A few of them are displayed in figure 4.3.

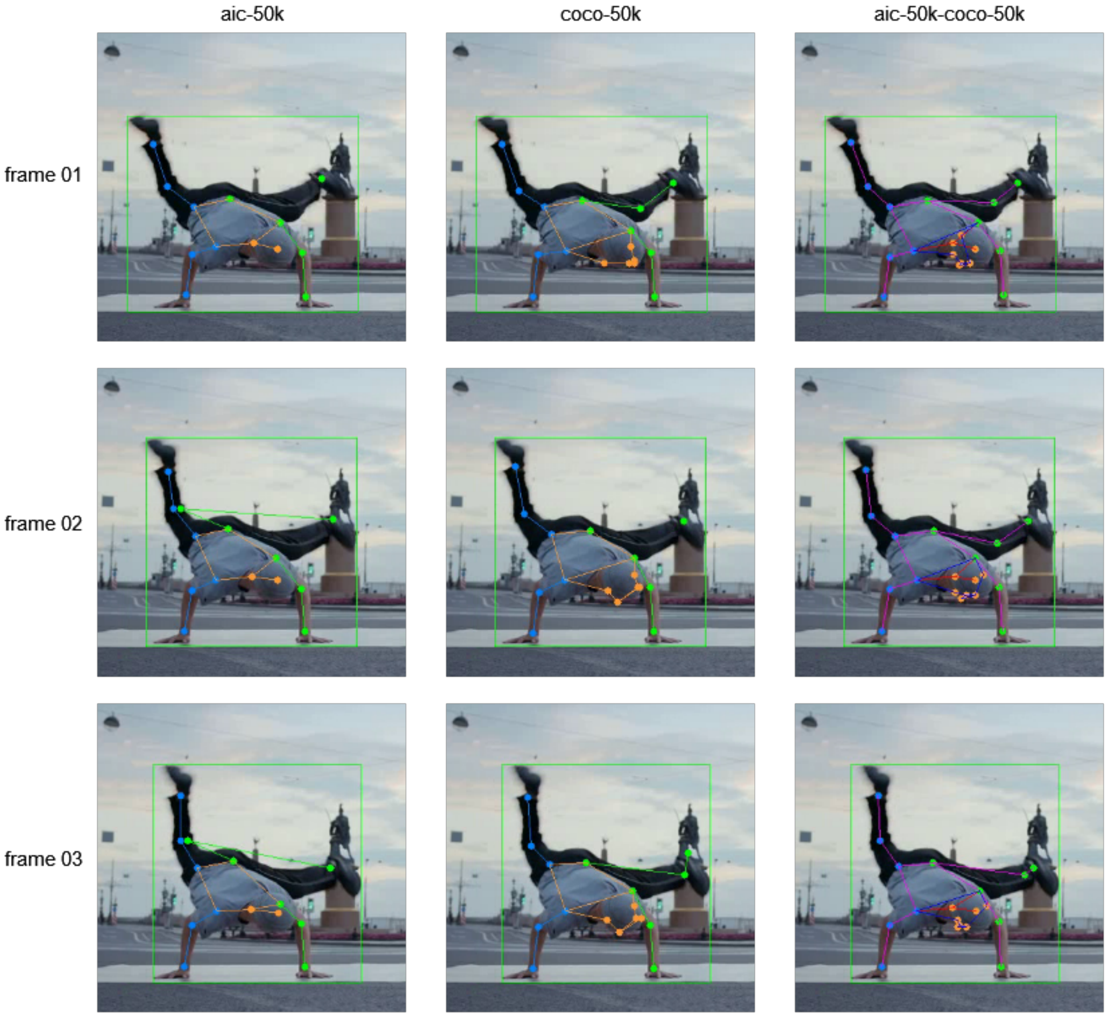


Figure 4.3: Example of frame-by-frame comparison, obtained by performing detection on each frame of a video. The composite model demonstrates a superior understanding of a difficult pose on these and more frames.

These results validate the previous findings in demonstrating superior pose detection by a composite-trained model on external data.

4.3 Discussion

Our findings demonstrate that models trained on composite datasets consistently outperform models trained on their parts when evaluating external data, by learning from their combined data. This establishes a general method of combining diverse datasets to enhance human pose detection in the wild. This work focuses solely on human pose detection, but the principle of allowing for dataset combination through a unified representation applies to other domains as well. The only caveat is that the datasets need to have keypoints in common.

Additionally, through a comparative analysis, we have investigated whether the observed improvements are solely attributed to the increase in dataset volume or if they truly stem from a reduction in selection bias. The composite-trained model outperforming single-dataset models trained on equivalent volume when evaluating on new data indicates that the performance enhancements achieved through the composite datasets are indeed a result of mitigating selection bias. Hence the composite dataset approach appears to effectively reduce the impact of dataset-specific biases, leading to more robust models that generalise better to new data.

Furthermore, we have examined the specific aspects of the data with which our models still struggle, aiming to identify valuable directions for future research. Our analysis reveals that rare and partly occluded poses pose a significant challenge for the current models, indicating the need for advancements in handling such scenarios. Addressing this challenge would further enhance the performance and robustness of human pose detection models in the wild.

Chapter 5

Conclusion

This work has presented a novel approach to human pose detection in the wild using composite datasets. By combining multiple datasets and mitigating the impact of selection bias, models trained on composite datasets consistently outperformed those trained on their parts. Our findings highlight the effectiveness of composite datasets in improving performance and generalisation capabilities in human pose detection tasks. The comparative analysis revealed that the observed improvements are at least partially a result of reducing selection bias rather than simply increasing dataset volume. Additionally, our analysis identified rare and partly occluded poses as an ongoing challenge for current models trained on popular datasets. The findings provide valuable insights for future research, demonstrating the potential benefit of a universal human pose representation and indicating the need for advancements in handling challenging scenarios. The successful implementation and evaluation of composite datasets will hopefully pave the way for further research on dataset combinations and contribute to the advancement of human pose detection for applications in the wild. Code, full results and documentation are available at `https://github.com/l-trochelmann/composite-training`.

Bibliography

[1] Jorge AC Ambrósio and Andrés Kecskeméthy. Multibody dynamics of biomechanical models for human motion via optimization. In *Multibody Dynamics: Computational Methods and Applications*, pages 245–272. Springer, 2007.

[2] Mykhaylo Andriluka, 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*, pages 3686–3693, 2014.

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[4] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose/tree/0.x`, accessed 18. 05. 2023, 2020.

[5] Haodong Duan, Kwan-Yee Lin, Sheng Jin, Wentao Liu, Chen Qian, and Wanli Ouyang. Trb: a novel triplet representation for understanding 2d human body. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9479–9488, 2019.

[6] Vittorio Ferrari, Manuel Marin-Jimenez, and Andrew Zisserman. Progressive search space reduction for human pose estimation. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[7] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[8] James J Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161, 1979.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[10] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander

Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

[11] John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. Mseg: A composite dataset for multi-domain semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2879–2888, 2020.

[12] 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.

[13] Tsung-Yi Lin, 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: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[14] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[15] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *Acm transactions on graphics (tog)*, 36(4):1–14, 2017.

[16] Thomas B Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding*, 104(2-3):90–126, 2006.

[17] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 483–499. Springer, 2016.

[18] Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. Strong appearance and expressive spatial models for human pose estimation. In *Proceedings of the IEEE international conference on Computer Vision*, pages 3487–3494, 2013.

[19] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019.

[20] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[21] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. Ai challenger: A large-scale dataset for going deeper in image understanding. *arXiv preprint arXiv:1711.06475*, 2017.

[22] Oliver Zendel, Katrin Honauer, Markus Murschitz, Daniel Steininger, and Gustavo Fernandez Dominguez. Wilddash-creating hazard-aware benchmarks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 402–416, 2018.

Erklärung der Urheberschaft

Hiermit versichere ich an Eides statt, dass ich die vorliegende Bachelor thesis im Studiengang Computer Science selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel - insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen – benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Ich versichere weiterhin, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe und die eingereichte schriftliche Fassung der auf dem elektronischen Speichermedium entspricht.

Ort, Datum Unterschrift

Hamburg, 22. 05. 2023

Erklärung zur Veröffentlichung

Ich stimme der Einstellung der Bachelor thesis in die Bibliothek des Fachbereichs Informatik zu.

Ort, Datum

Hamburg, 22. 05. 2023

Unterschrift