Mission Im-Poseable

Tom Byrd byrd0134@umn.edu
Alex Jonas jonas060@d.umn.edu
David Kong kong0111@umn.edu
Mason Zarns zarns008@umn.edu

Abstract

Established computer vision human-pose-estimation techniques generalize poorly to non-human primates. The OpenMonkeyChallenge is a competitive pose estimation benchmark that provides an extensive data set of annotated non-human-primate images, enabling detailed comparison between experimental pose-estimation techniques [27]. The paper presented details two proposed architectures compared with an established baseline architecture - the Convolutional Pose Machine (CPM) [25]. The baseline architecture outperformed the proposed Partial U-Net [2] and Modified CPM architectures in nearly every metric.

1. Introduction

The global market size of computer vision applications is projected to grow to over $41 billion by 2030. [1] The explosive growth in the field over the last decade was catalyzed by the incorporation of deep learning into methodological approaches, beginning with the first ImageNet Large-Scale Visual Recognition Challenge in 2012. [6] Since then researchers and industry partners have discovered countless ways to apply computer vision to solve some of societies most pressing challenges.

One area of considerable focus is the detection of living organisms across space and time. The goal is to quantify and therefore be able to predict and analyze physiological positions. The goal of physiological prediction extends beyond simply placing a bounding box around an organism on a screen. Instead we wish to recognize *how* a person is positioned and moving within that box, i.e., their *POSE*. By capturing poses, we can capture their complicated nuances, which may include physical function, behavioral intent/body language, and health.

The majority of work on POSE RECOGNITION using computer vision has been done on humans, owing to the large corpus of readily available human images. Initial deep architectures like DeepPose [24] and the Convolutional Pose Machine (CPM) [25] inspired more recent architectures like HRNet [23] which have been able to show impressive performance on human subjects. Within the past few years an emergence of pose recognition methods for non-human primates has gained footing within academia.

Pose recognition in animals is arguably as important as pose recognition in humans. Animal studies form the bedrock of basic science laboratories and are crucial in the development and clinical trial processes for life-saving pharmaceuticals. From fruit flies to zebra fish to chimpanzees, laboratory animals are ubiquitous within academic research. Our ability to accelerate and translate scientific discoveries from animals to the human population depends on our ability to intimately understand the animals behavior. Pose estimation aids us understanding the animals behavior.

Non-human primates (NHP) like chimpanzees are particularly important for the development of generalizable pose recognition software. NHPs share the closest similarity to humans in terms of behavior and pharmacodynamics. Due to the similarity between NHPs and humans laboratory studies benefit from pose estimation as researchers are able to tease out changes to NHPs brought on by experimental conditions.

Beyond the lab NHPs are often the focus of environmental preservation. Pose Estimation supports the tracking and monitoring of communities of NHPs in nature.

NHP pose estimators will be able to improve our ability to recognize atypical poses in humans. Elderly patients would benefit from a video system that monitors and detects falls for independent seniors living alone in a community. [6] A fallen human will have a pose that is horizontal and likely less familiar to a human-trained pose recognizer.

The challenges to pose estimation in NHPs stem from NHP poses being quite different from the corpus of standard human poses (e.g. horizontal on a tree). Thus traditional human-trained pose estimators do not perform well on NHP data. [27].

Here, our aim is to improve NHP pose estimation using a publicly available repository of NHP images, contained in the OpenMonkeyChallenge. [27] The data will be utilized to train a baseline pose estimator; train two novel pose estimators; compare the results of the novel pose estimators against the baseline method.

The data consist of 111,529 images of 26 NPH species, collected from the internet, primate research centers, and

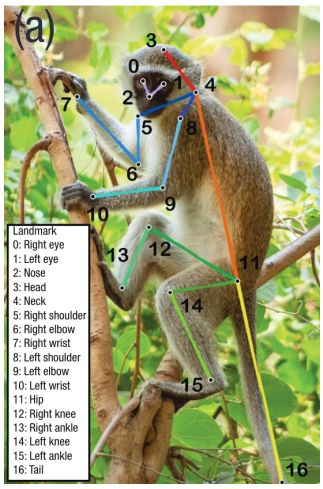

Figure 1. Landmarks for non-human primate pose estimation.

the Minnesota Zoo. There are 66,917 images in the training set, 22,306 in the validation set, and 22,306 in the testing set. Each image is labeled with 17 landmarks (key points) that comprise the overall pose (Figure 1). We will measure our performance using three metrics: mean per joint position error (MPJPE); probability of correct keypoint at error tolerance (PCK) with error tolerance of 0.5; and average precision based on object keypoint similarity (AP).

2. Related Work

The challenge of pose estimation is a thoroughly researched problem particularly for human pose estimation. Naturally there are a vast number of existing strategies to address the problem of pose estimation. All notable models today are varying architectures that all utilizing Convolutional Neural Networks [23].

There are two main groups of architectures for human and non-human pose estimation, Top-down and Bottom-up. In the former, the model first identifies a bounding box with an entity inside. Then, performs single-person pose estimation on the cropped image of the bounding box, as described in Section 2.2. Section 2.1, details modern architectures associated with the the latter, Bottom-up approach. The bottom-up approach detects all possible body parts in an image individually. The bottom-up model then attempts to group the individual body parts into recognizable skeletons. Either of the two approaches could potentially be expanded to detect multiple skeletons in a single image. Some modern techniques allow for real-time pose estimation from video input, at varying frame rates. A common architecture, used to compare with newer models is SimpleBaseline [26].

2.1. Bottom-up

An advantage of bottom-up architectures is the potential for generalization. In 2016, Pischulin *et al* proposed DeepCut [19]. Which detected landmarks before labeling, filtering, and assembling the parts into skeletons. An improved model was proposed by Insafutdinov *et al*, DeeperCut [9], built around ResNet, a common standard for image feature extraction [27]. DeeperCut provided a more sophisticated, incremental strategy to the assembly of skeletons phase. Two years later, DeepLabCut [13] was developed to generalize for animal pose estimation with great success. DeepLabCut is a relatively shallow model with fewer parameters than many competing models [10]. A variant of ResNets are trained on the annotated training images, then the outputs are passed into deconvolutional layers that predict spatial densities for each feature. The advantages of the bottom-up approach is the generalizability the method provides. Many animal types have been estimated successfully utilizing the DeepLabCut bottom-up approach.

Osokin *et al* proposed OpenPose [17] in 2018, introducing Part Affinity Fields to decompose the matching phase into more systematically addressable mathematical representations. Despite the advantages, Bottom-up approaches struggle to solve the occlusion problem, when parts of bodies are unseen or multiple bodies are overlapping in an image [3].

2.2. Top-down

The Top-down approach requires a method for detecting human bodies and generating corresponding bounding boxes. Once these are generated, each body can be estimated from a more predictable cropped image [16]. AlexNet represented a significant breakthrough in the vision community, sparking a myriad of interest and research [12].

A very recent and promising architecture, AlphaPose, can perform accurate whole-body pose estimation and for multiple bodies in real time [7]. Symmetric Integral Keypoint Regression improves local feature extraction and Parametric Pose Non-Maximum-Suppression assists in eliminating redundant attributes [14].

The established HRNet architecture [28] has been modified using a variety of strategies to make it more lightweight in terms of training complexity and generalizability, with great success [11].

The medical community has a vested interest in the advancement of classification architectures capable of predicting diagnoses [2]. Celebi *et al* propose Cancer-Net, specifically designed for the purpose of identifying malignant melanomas [4]. Du *et al* propose the UNet architecture in 2020 for medical image segmentation [5]. Zhang *et al* proposes techniques and guidelines for deployment of computer vision technologies to modern medicine practices [29].

Zhang *et al* demonstrate that allowing an architecture room to "forget" aspects of what it has learned during training can afford advantages in generalizability and preventing overfitting [30].

The OpenMonkeyChallenge is one of many datasets like the ImageNet Database of annotated images for the purpose of comparing modern convolutional neural network architectures for the problem of pose estimation experimentation [8].

3. Objective Functions

The baseline and proposed architectures utilize 3 objective functions measuring key point estimation.

$$MPJPE_i = \frac{1}{J}\sum_{j=1}^{J}\frac{\|\hat{x}_{ij} - x_{ij}\|}{W} \tag{1}$$

Equation 1 known as the *Mean per joint position error (MPJPE)* measures the normalized error between the predicted and ground truth for each landmark. The smaller the overall loss value the better.

$$PCK@\epsilon = \frac{1}{17J}\sum_{j=1}^{J}\sum_{i=1}^{17}\delta\left(\frac{\|\hat{x}_{ij} - x_{ij}\|}{W} < \epsilon\right) \tag{2}$$

Equation 2 known as *probability of correct keypoint (PCK)* is defined by the prediction accuracy given error tolerance (error tolerance 0.5). The larger the loss value the better.

$$AP@\epsilon = \frac{1}{17J}\sum_{j=1}^{J}\sum_{i=1}^{17}\delta\left(OKS_{ij} \geq \epsilon\right) \tag{3}$$

Equation 3 is our *Average Precision* measuring prediction precision. The larger the loss value the better.

$$OKS_{ij} = exp\left(-\frac{\|\hat{x}_{ij} - x_{ij}\|^2}{2W^2k_i^2}\right) \tag{4}$$

Equation 4 a subset of Equation 3 defined as our *Key point similarity measurement*. Measures per landmark accuracy by taking into account per landmark variance e.g. an eye is visually less ambiguous then ears.

4. Baseline Method

Historically pose estimation was limited to human pose estimation. Two sources of complexity were quickly discovered during initial estimation of human poses. The first being the large number of degrees of freedom 20 of the underlying skeleton leading to a high dimensional configuration space to search over. The second being variation of appearance of people in images. Methods for estimating articulated poses generally of a human from an image had revolved around graphical models. The graphical models would capture the correlations and dependencies between the locations of the parts [21].

The graphical models however contained a few major drawbacks. Inference in graphical models were difficult to learn and often inexact in all but a few models; namely the tree-structured and star-structured models. Despite the tree and star structured models ability to learn inference they were unable to capture important dependencies. The important dependencies being dependencies between locations of each of the parts which led to characteristic errors. A notable characteristic error being double counting. Meaning the same part of the image is used to explain more then one part. Double counting often occurs due to the symmetrical appearance of the body (e.g. the left and right knee are similar in appearance). To solve the issue of symmetrical appearance and other characteristic errors such as self-occlusion required the use of approximate inference making parameter learning difficult [21].

The second problem with graphical models is defining potential functions. The function selection is usually of a parametric form to allow traceability inference. In order to establish efficient inference most approaches are restricted to using simple classifiers for part detection. The choices are guided by tractability of inference rather then the complexity of the data. The result is a restricted model that can not address the overall complexity of the problem presented.

4.1. Convolutional Pose Machine

The Convolutional Pose Machine avoids the restriction imposed by graphical models by directly training the inference procedure rather then the model being guided by tractability of inference.

Leveraging the usage of Convolutional Neural Networks and *Pose Machine* architecture the Convolutional Pose Machine allows us to learn spatial models of the relationships between parts. The Convolutional Pose Machine has surpassed benchmark performance on MPII, LSP, and FLIC datasets.

In order to tease out exactly what is a Convolutional Pose Machine we will break the architecture down into its main components. The main components are the *Pose Machine* architecture and Convolutional Neural Network Architecture.

4.2. Pose Machine

The *Pose Machine* architecture builds off the hierarchical inference machine used for scene parsing.

The Pose Machine is a sequential prediction algorithm which emulates the mechanics of message passing to predict confidence for each part. Improving its estimates with

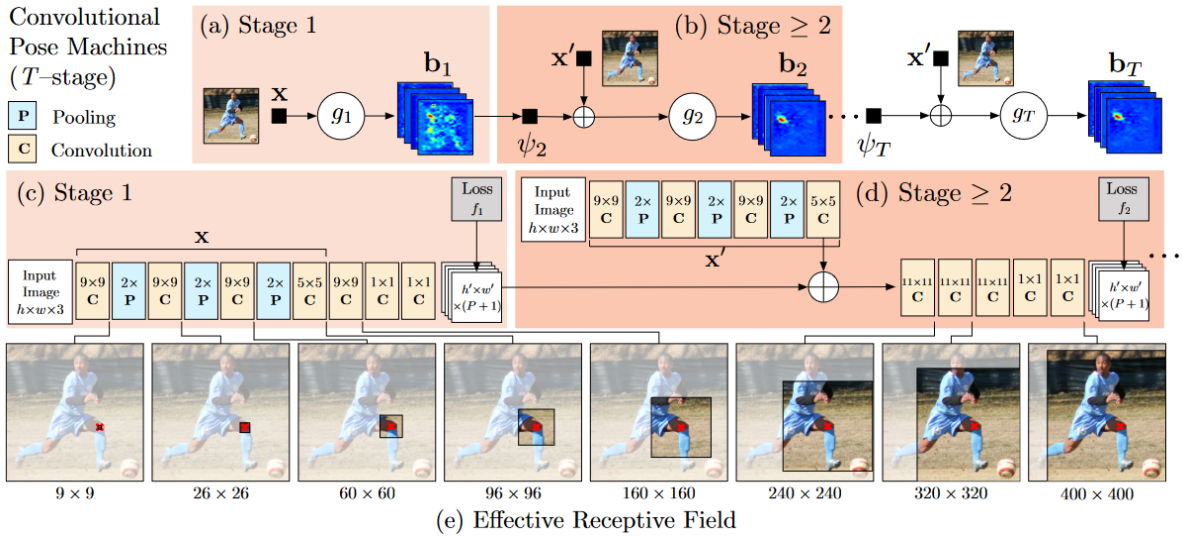

Figure 2. Convolutional Neural Network

each stages iteration. [21]. The Pose machine reduces double counting by incorporating more interactions among multiple variables at a time. The Pose machine also learns a spatial model directly from the data without needing to specify the parametric form of potential functions thereby not falling into the restrictive trap of graphical models. The modular architecture of the Pose Machine allows the usage of high capacity predictors which are beneficial for multi-modal appearance of each part (a knee may be facing forward or to the side within a given image).

The pose machine also incorporates multi-scale cues within its architecture by modeling the hierarchy of parts. The multi-scale cues [25] allow conditioning of finer part detection on the detection of larger parts improving localization with the result of an improved accuracy of part detection.

4.3. Convolutional Neural Network Motivation

The motivation behind the usage of Convolutional Neural Networks is the ability to learn spatial dependencies generated from belief maps without the need for carefully made priors, careful initialization or dedicated precision refinement. [25]

The spatial dependencies can then be used to refine the belief maps at each stage of the sequential prediction framework. Meaning we are able to encode spatial relationships without priors or a parametric form. The relationships are then fine-tuned through iteration resulting in increased accuracy on prediction of parts.

4.4. Baseline Approach

The baseline implementation will be leveraged from the the published paper *Convolutional Poses* as published by Shin-En Wei et al. through *The Robotics Institute Carnegie Mellon University* [25].

Figure 2 details the convolutional architecture and receptive fields across layers for a Convolution Pose Machine with T stages. The Pose Machine is detailed in sections (a) and (b) of figure 2 while the corresponding convolutional networks are detailed in sections (c) and (d). Sections (a) and (c) show the architecture that operates only on image evidence from the first stage. While sections (b) and (d) illustrate the architecture repeated for all subsequent stages of 2 through T. Our baseline approach utilized 6 stages in total one initial and 5 subsequent layers.

The overall architecture is locally supervised after each stage with an intermediate loss layer to prevent vanishing gradients during training.

Section (e) illustrates the effective receptive field on an image that is centered at the left knee. The large receptive field allows the model to capture "long-range" spatial dependencies such as the head and knee of the image.

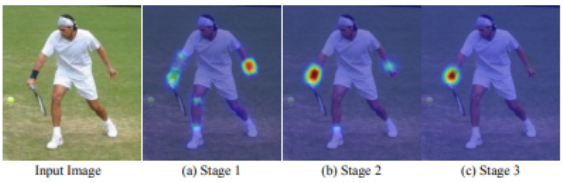

Figure 3. Belief Map

Figure 3 illustrates increasingly precise belief map predictions for the location of the *right elbow* in each stage as defined in figure 2.

5. Proposed Architectures

5.1. Modified U-Net, a.k.a. SlashNet

U-Net is a convolutional neural network (CNN) architecture developed for biomedical image segmentation. It was first proposed by Olaf Ronneberger, Philipp Fischer, and Thomas Brox in 2015. [22] The U-Net architecture is designed to perform image segmentation tasks by learning to localize and classify different objects in an image. It consists of a contracting path and an expanding path, which are connected by a series of skip connections. The contracting path is responsible for extracting features from the input image, while the expanding path is responsible for mapping the features back to the original image size and generating a segmentation map. One of the key features of U-Net is the use of skip connections, which allow the network to incorporate both high-level and low-level features from the input image. These skip connections help to preserve spatial information and enable the network to learn more complex and nuanced image features. U-Net has been widely used in a variety of image segmentation tasks, particularly in the biomedical field, and has achieved state-of-the-art results in many benchmarks. [2] It has also been applied to other areas such as satellite image segmentation and street scene segmentation.

However, predicting target heatmaps (as in the case of pose estimation) poses a challenge to the U-Net architecture. Heatmaps are sparse, with only a few value pixels surrounded by a sea of non-target pixels. Because of this sparsity, CNNs have trouble detecting and converging on a high resolution heatmap solution. To increase the ratio of signal to noise, we decrease the size of our heatmap labels to 32 by 32 pixels, which allows algorithms like CPM to detect the signal. However, the baseline U-Net output dimensions are the same size as the input image, in our case 256 by 256 pixels. We hypothesized that a high resolution output heatmap could not be solved by the traditional U-Net architecture. Instead, we proposed SlashNet, a modification of U-Net where the upconvolution laters and associated concatenations are "slashed" away, leaving only one upconvolution and concantenation near the base of the U, as depected in 4. We reasoned that the high channel volume at this lower level would be sufficient to preserve semantic segmentation of the lower resolution heatmap and would lead to faster and more accurate convergence.

The second major modification we made to our SlashNet was our loss function. Traditionally, mean squared error has been used as the loss function in pose recognition algorithm training. This loss function suffers from the same sparsity of heatmap projections: incorrect pixel-wise predictions over the target (joint) pixels will influence the loss but will be out weighted by the majority of pixels in the majority class that are closely aligned. Instead, we used Intersection-over-Union (IoU) loss as our loss function to train SlashNet.

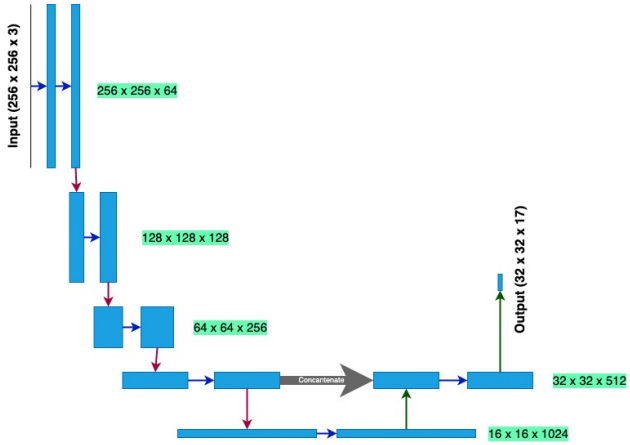

Figure 4. SlashNet architecture

IoU loss calculates the overlap between the prediction and the ground-truth images, and it is calculated by dividing the size of the intersection between the two by the union of the two. IoU in theory should outperform a loss function like MSE because if can account and correct for class imbalance, as in the case of sparse heatmaps. [20]

5.2. Upscaled CPM

A limitation within our baseline method [25] was the use of three 2x2 max pooling layers resulting in heat maps that were downscaled eight times. The downscaling was a compromise that allowed the Baseline CPM to achieve a large receptive field while sacrificing a degree of precision in order to reduce computational power needed. We believed we could reclaim some of the loss of precision at the expense of computational load. It is suggested in [25] that a large receptive field could be achieved with pooling layers, increased kernel sizes, or more convolutional layers. It was clear that the receptive field from the max pooling layers is effective in learning the relationship between parts. Therefore we hypothesized we could easily achieve a higher resolution through upsampling, found in many related works [13, 15, 22, 26]. The use of the skip step [22] was introduced as a simple way to propagate high resolution data from our input image to the upscaled heatmap. The goal of the modified CPM algorithm was to keep the network as simple as possible; improving precision while keeping training times reasonable. To achieve the aim of simplicity, improved precision and maintaining a similar computational load to the baseline method we utilized a upsampling layer in the form of a transposed convolution. Along with incorporating a skip step since additional layers would result in a dramatic increase in computational load during training.

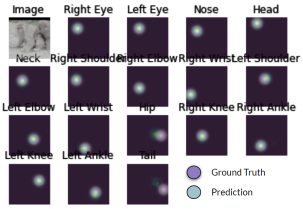

Figure 5. Baseline CPM joint predictions vs. ground-truth

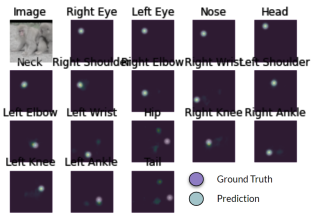

Figure 6. Modified CPM joint predictions vs. ground-truth

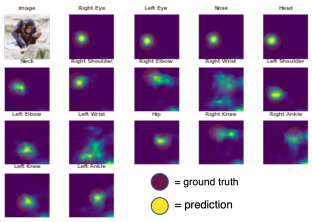

Figure 7. SlashNet joint predictions vs. ground-truth

6. Results

6.1. Prediction Heatmap Comparison

SlashNet utilizing Intersection of Union loss was able to predict joint locations with moderate ability 7. Primate's joints when obscured or close to external objects were difficult for the SlashNet to accurately identify. Training speed of SlashNet was considerably faster than both our basline and proposed CPM frameworks. The increased training speed of SlashNet was due to the removal of nearly half the Up-Convolution Layers. The lack of resolution reconstitution that resulted from the removal of the Up-Convolution Layers limited the ability of SlashNet to fully understand spatial relationships among key points resulting in an increase an erroneous predictions.

6.2. Objective Function Performance

Metric	Base CPM	SlashNet	Mod. CPM
Mean MPJPE	0.0664	0.253	0.0631
PCK@0.5	0.9918	0.805	0.9881
AP	0.9949	0.831	0.9925

6.3. Challenges in Implementation

The development and implementation of our models were cut shorter than preferred as we began to run out of time. Errors and growing pains from learning to use new tools and frameworks (such as Colab and PyTorch) caused our initial models to train very slowly and made debugging difficult with slow results. Our initial iteration of our baseline model was not training as expected, leading to changes in its implementation in PyTorch. Initially, the baseline CPM would use the same `Conv2d` filters and weights for each stage after the first, but was changed to have separate values for every stage. Due to oversight and time restraints, our modified, upscaled CPM was not trained with these same implementation changes. The result was that our baseline CPM achieved the results we were expecting, but our modified CPM achieved very similar results while converging to a saturated result quicker than expected. Both trained for roughly the same amount of time, with the baseline training for around ten hours and the modified CPM a few hours longer, while making seemingly little improvement towards performance. We believe that the same implementation change in the modified CPM would result in a notable improvement in precision compared to the baseline with a significant cost of computational complexity and training time.

7. Future Work

Potential future improvements include "strategic forgetting" or *active forgetting* while leveraging the CPM framework. Memory and forgetting mechanisms are interwoven within humanities learning and memory cognition. Therefore it would appear beneficial to replicate the evolutionary biological strategy of learning - that is forgetting strategically. Future work will utilize an "active forgetting mechanism" via a "plug-and-play" forgetting layer [18]. The idea would be to utilize the forgetting layer after each CPM Stage prediction. The intended outcome being the forgetting of noisy pixels with emphasis placed on pixels of importance.

8. Conclusion

Given two novel implementations for pose estimation our team succeeded in the intended goal of improving NHP Pose Estimation given a robust baseline method.

The proposed methods of SlashNet (MPJPE: 0253; PCK: 0.805; AP: 0.831) and our Modified CPM (MPJPE: 0.0631; PCK: 0.9881; AP :0.9925) failed to attain the same level accuracy in almost every metric as our Baseline CPM implementation (MPJPE 0.0664; PCK: 0.9918; AP: 0.9949). The only exception was our modified CPM model was able to achieve a slightly lower MPJPE metric value 0.0631 compared to 0.0664 - a difference of 0.0033.

The metric of MPJPE should not be overlooked. MPJPE (Mean per point positional error) measures how far the models prediction of the joint was from ground truth. Our novel CPM method was able to achieve lower per joint positional error then our basline method. Improving NHP Pose Estimation over the baseline method.

9. GitHub Link

https://github.com/zarns/openMonkeyChallenge

References

[1] N Abhijith. Computer vision market predicted to attain $41.11 billion by 2030. 1

[2] Vatsala Anand, Sheifali Gupta, Deepika Koundal, Soumya Ranjan Nayak, Paolo Barsocchi, and Akash Kumar Bhoi. Modified u-NET architecture for segmentation of skin lesion. 22(3):867. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute. 1, 2, 5

[3] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. 2

[4] M. Emre Celebi, Quan Wen, Hitoshi Iyatomi, Kouhei Shimizu, Huiyu Zhou, and Gerald Schaefer. A state-of-the-art survey on lesion border detection in dermoscopy images. pages 97–129. 2

[5] Getao Du, Xu Cao, Jimin Liang, Xueli Chen, and Yonghua Zhan. Medical image segmentation based on u-net: A review. 64(2):20508–1–20508–12. 2

[6] Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean, and Richard Socher. Deep learning-enabled medical computer vision. 4(1):1–9. Number: 1 Publisher: Nature Publishing Group. 1

[7] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. AlphaPose: Whole-body regional multi-person pose estimation and tracking in real-time. 2

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. ISSN: 1063-6919. 3

[9] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. DeeperCut: A deeper, stronger, and faster multi-person pose estimation model. 2

[10] Jessy Lauer, Mu Zhou, Shaokai Ye, William Menegas, Steffen Schneider, Tanmay Nath, Mohammed Mostafizur Rahman, Valentina Di Santo, Daniel Soberanes, Guoping Feng, Venkatesh N. Murthy, George Lauder, Catherine Dulac, Mackenzie Weygandt Mathis, and Alexander Mathis. Multi-animal pose estimation, identification and tracking with DeepLabCut. 19(4):496–504. 2

[11] Shiqi Li and Xiang Xiang. Lightweight human pose estimation using heatmap-weighting loss. 2

[12] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. 2

[13] Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. 21(9):1281–1289. Number: 9 Publisher: Nature Publishing Group. 2, 5

[14] Laurie Needham, Murray Evans, Darren P. Cosker, Logan Wade, Polly M. McGuigan, James L. Bilzon, and Steffi L. Colyer. The accuracy of several pose estimation methods for 3d joint centre localisation. 11(1):20673. 2

[15] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. 5

[16] Elisha Odemakinde. Human pose estimation with deep learning - ultimate overview in 2022. 2

[17] Daniil Osokin. Real-time 2d multi-person pose estimation on CPU: Lightweight OpenPose. 2

[18] Jian Peng, Xian Sun, Min Deng, Chao Tao, Bo Tang, Wenbo Li, Guohua Wu, QingZhu, Yu Liu, Tao Lin, and Haifeng Li. Learning by active forgetting for neural networks. 6

[19] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. DeepCut: Joint subset partition and labeling for multi person pose estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4929–4937. IEEE. 2

[20] Md Atiqur Rahman and Yang Wang. Optimizing intersection-over-union in deep neural networks for image segmentation. In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Fatih Porikli, Sandra Skaff, Alireza Entezari, Jianyuan Min, Daisuke Iwai, Amela Sadagic, Carlos Scheidegger, and Tobias Isenberg, editors, *Advances in Visual Computing*, volume 10072, pages 234–244. Springer International Publishing. Series Title: Lecture Notes in Computer Science. 5

[21] Varun Ramakrishna, Daniel Munoz, Martial Hebert, James Andrew Bagnell, and Yaser Sheikh. Pose machines: Articulated pose estimation via inference machines. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, volume 8690, pages 33–47. Springer International Publishing. Series Title: Lecture Notes in Computer Science. 3, 4

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. 5

[23] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5686–5696. ISSN: 2575-7075. 1, 2

[24] Alexander Toshev and Christian Szegedy. DeepPose: Human pose estimation via deep neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1653–1660. 1

[25] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. 1, 4, 5

[26] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In Vittorio Ferrari,

Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, volume 11210, pages 472–487. Springer International Publishing. Series Title: Lecture Notes in Computer Science. 2, 5

[27] Yuan Yao, Abhiraj Mohan, Eliza Bliss-Moreau, Kristine Coleman, Sienna M. Freeman, Christopher J. Machado, Jessica Raper, Jan Zimmermann, Benjamin Y. Hayden, and Hyun Soo Park. OpenMonkeyChallenge: Dataset and benchmark challenges for pose tracking of non-human primates. page 2021.09.08.459549. 1, 2

[28] Changqian Yu, Bin Xiao, Changxin Gao, Lu Yuan, Lei Zhang, Nong Sang, and Jingdong Wang. Lite-HRNet: A lightweight high-resolution network. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10435–10445. IEEE. 2

[29] Angela Zhang, Lei Xing, James Zou, and Joseph C. Wu. Shifting machine learning for healthcare from development to deployment and from models to data. pages 1–16. Publisher: Nature Publishing Group. 2

[30] Baosheng Zhang, Yuchen Guo, Yipeng Li, Yuwei He, Haoqian Wang, and Qionghai Dai. Memory recall: A simple neural network training framework against catastrophic forgetting. 33(5):2010–2022. Conference Name: IEEE Transactions on Neural Networks and Learning Systems. 3