Project Report - ECE 285
Unipose with DARKPose

Lulua Rakla
Department of Electrical and Computer Engineering
A59012895

Shanmukha Vellamcheti
Department of Electrical and Computer Engineering
A59005427

Abstract

Human Pose Estimation is a widely researched problem in computer vision. The objective is to find the joint locations of the human body such as shoulder, hip, knee etc. from images. Pose Detection is used in a number of downstream tasks such as activity recognition, surveillance, fitness training, augmented reality etc. In most cases, pose detection is done in two stages - first detecting the human and then regressing to keypoint locations. This makes the pose detection pipeline slow. Hence, we decided to implement Unipose, a pose detection algorithm which does keypoint and bounding box detection in one stage. The algorithm uses a WASP module (Waterfall Atrous Spatial Pooling) which uses dilated convolution to increase FOV. Another problem in pose estimation is that the heatmap is at a lower resolution than the image. To recover the coordinate location in image space, DARK (Distribution-Aware coordinate Representation of Keypoint) uses efficient Taylor-based coordinate decoding and unbiased coordinate encoding. In this project, we aim to integrate DARKPose with Unipose and test if coordinate encoding and decoding improves the performance.

1 Introduction

Pose estimation is a challenging computer vision problem. The challenge stems from the large number of degrees of freedom in the human body mechanics and the frequent occurrence of parts occlusion. To overcome problems with occlusion, many methods rely on statistical and geometric models to estimate occluded joints. The two algorithms - UniPose [1] and DARKPose [2] that we have chosen address two big issues in computer vision.

UniPose incorporates contextual segmentation and joint localization to estimate the human pose in a single stage, with high accuracy, without relying on statistical postprocessing methods. The Waterfall Atrous Spatial Pooling (WASP) in UniPose leverages the efficiency of progressive filtering in the cascade architecture, while maintaining multi-scale fields-of-view (FOV) comparable to spatial pyramid configurations. With UniPose-LSTM, the algorithm gives SOTA results on pose detection in videos as well.(This is out of scope for this iteration of the project).

DARK(Distribution-Aware coordinate Representation of Keypoint) Pose refines the coordinate encoding and decoding process of heatmaps - the most widely used joint representation format. Heatmap is characterised by giving spatial support around the groundtruth joint locations, considering not only the contextual clues but also the inherent target position ambiguity. An obstacle with heatmaps is that the computational cost is quadratic function of the input image resolution, requiring input images to be downsampled. To predict joints in the original image coordinate space, after the

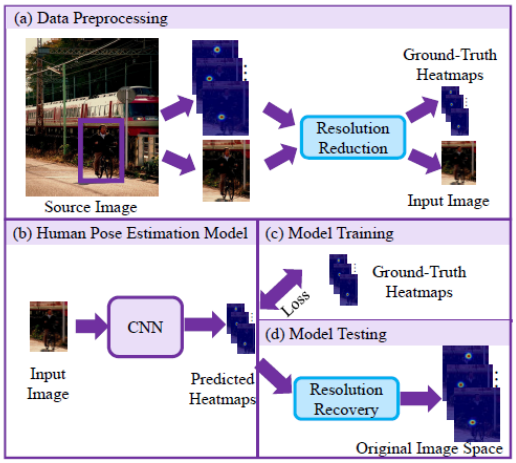

Figure 1: For efficiency, resolution reduction is often applied on the original person detection bounding boxes as well as the groundtruth heatmap supervision. That is, the model operates in a low-resolution image space. At test time, a corresponding resolution recovery is therefore necessary in order to obtain the joint coordinate prediction in the original image space.

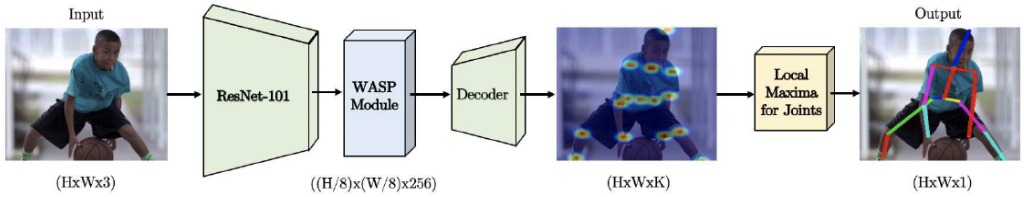

Figure 2: UniPose architecture for single frame pose detection. The input color image of dimensions (HxW) is fed through the ResNet-101 backbone and WASP module to obtain 256 feature channels at reduced resolution by a factor of 8. The decoder module generates K heatmaps, one per joint, at the original resolution, and the locations of the joints are determined by a local max operation.

heatmap prediction requires resolution recovery. This is called coordinate decoding. DARKPose is a distribution-aware coordinate representation method for more accurate joint localisation at sub-pixel accuracy. It accounts for the distribution information of heatmap activation via Taylor-expansion based distribution approximation. Besides, ground-truth heatmaps suffers from quantisation errors when being generated, leading to imprecise supervision signals. Therefore DARKPose uses unbiased heatmaps allowing Gaussian kernel being centred at sub-pixel locations. In the next sections we aim to qualitatively and quantitatively prove if adding DARK heatmap encoding and decoding improves the performance of keypoint human pose estimation using UniPose. We also perform some ablation studies like 1) using pretrained weights and fine-tuning with and without DARK 2) Coordinate encoding vs decoding

2 Related Work

2.1 UniPose

Traditionally pose estimation involved estimating the joints using the geometry between joints. In the Deep Learning era, Convolution Neural Networks have achieved superior results. Some of the most popular pose detection algorithms are OpenPose [7] which uses Part Affinity Fields. PAF uses the detection of more significant joints to better estimate the prediction of less significant joints. This innovation allowed advances toward multi-person detection with decreased complexity and computational power.The High-Resolution Network (HRNet) [8] includes both high and low resolution representations. Starting with high resolution, the method gradually adds low resolution

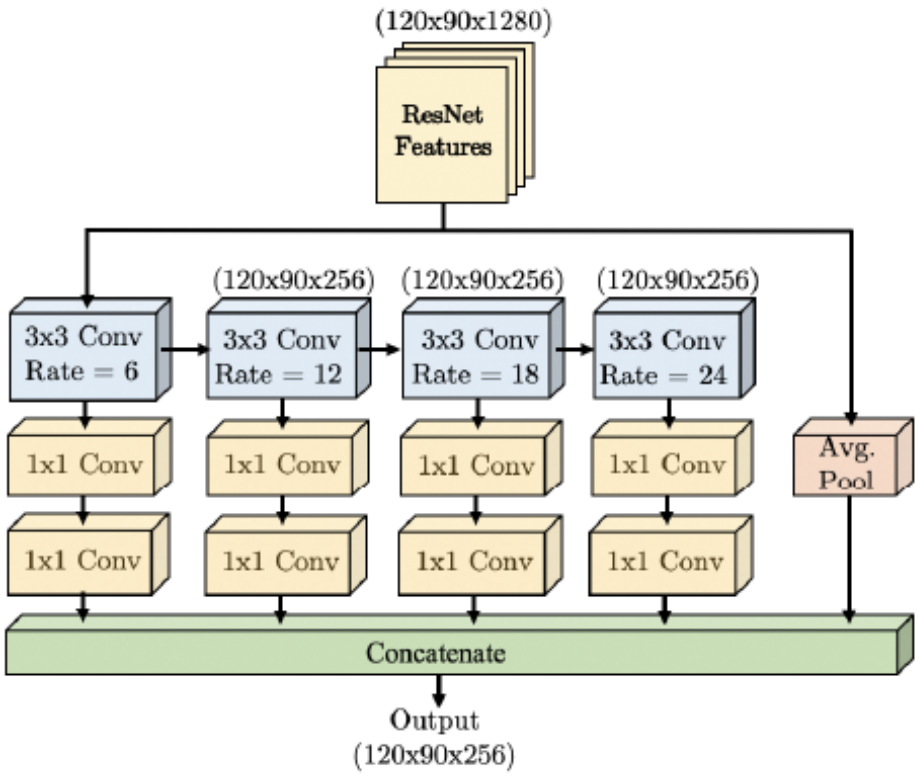

Figure 3: Architecture of WASP module. The inputs to the WASPmodule are 1280 channels of ResNet-101 features maps.

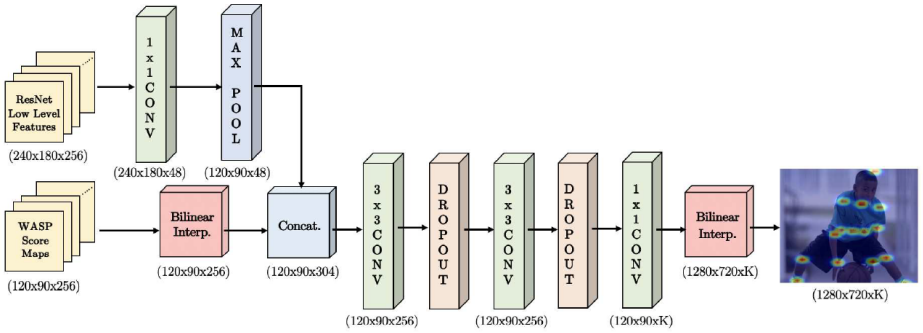

Figure 4: Architecture of Decoder module. The inputs to the decoder are 256 channels of ResNet low level features and 256 channels of theWASP feature maps. The output of the decoder is K heatmaps corresponding to K joints, shown in the image example.

sub-networks to form more stages, and performs multi-scale fusion between sub-networks. HRNet benefits from the larger FOV of multi resolution, a capability that is achieved in a simpler way with WASP model of UniPose.A drawback of some methods is that they require an independent branch for the detection of the bounding box of human subjects in the frame e.g LightTrack [10]and LCR-Net [9].

2.2 DARKPose

Only a handful of methods directly regress to joint coordinates without heatmap. e.g [12] This form of representation lacks spatial and contexual information, making the learning of joints very challenging. Heatmaps were introduced in [13] and became most popular form of coordinate representation. Generally, the mainstream research focus is on designing network architectures for more effectively regressing the heatmap supervision. In contrast to previous works, DARKPose targets the issue of heatmap representation on human pose estimation.It involves a principled coordinate representation method for significantly improving the performance of existing models. It is a method which can be seamlessly integrated in any existing model without any design changes.

3 Method

3.1 Unipose

The pipeline we followed for Unipose is shown in Fig 2. The input image's features are extracted using ResNet-101. The final layers of ResNet are replaced by the WASP module. The resultant feature maps are processed by the decoder module. The decoder generates K heatmaps, one for each joint. The probability distributions of joints are obtained using a softmax. The original resolution is recovered by bilinear interpolation, followed by local max operation to localize the joints.

3.1.1 WASP Module

The WASP architecture is shown in Fig. 3. WASP uses Atrous convolutions (dilated convolutions) to increase FOV. It also has a cascade of atrous convolutions at increasing rates. It has a waterfall structure, where the features are processed through a filter and then parallelized. To achieve multi-sclae representation, streams of all brances, and avg pooling of input is combined.

3.1.2 Decoder Module

The decoder module converts the score maps to heatmaps. The architecture is shown in Fig. 4
To get started, we implemented the baseline model of UniPose as outlined in the paper. The first step was to get the ground truth heatmaps with gaussian kernels from the raw coordinate information. The heatmaps were downsampled by a stride of 8. After that, we designed the training and validation loops. Training was done with batch size of 8 using Adam Optimizer and MSELoss. We finetuned the dataset for 10 epochs using the pretrained weights. We also ran trained the model from scratch for 20 epochs. The performance of the model was not equal to what was reported in the paper as the model was trained on the LSP-extended dataset (10k images) whereas here we only used the basic LSP dataset (2k images) due to resource constraints. To decrease the iteration time of the model, we increased the validation batch size to 10 from 1.

3.2 DARKPose

3.2.1 Encoder module

In order to gain the contextual information from the images proper encoding procedure has to be followed. The usual convention is to convert the keypoints to heatmaps using a Gaussian representation of the coordinates. This method also downsamples the original image into the model input size. This scaling factor is usually called 'stride'. Usually the scaled coordinates are rounded off to the nearest integer but in DARK the precision remains intact. This will allow for less biased representation of coordinate locations thereby giving improved ground truth heatmaps. This entire process is summarised via the equations below,

$$g^{'} = (u^{'}, v^{'}) = \frac{g}{\lambda} = (\frac{u}{\lambda}, \frac{v}{\lambda}) \tag{1}$$

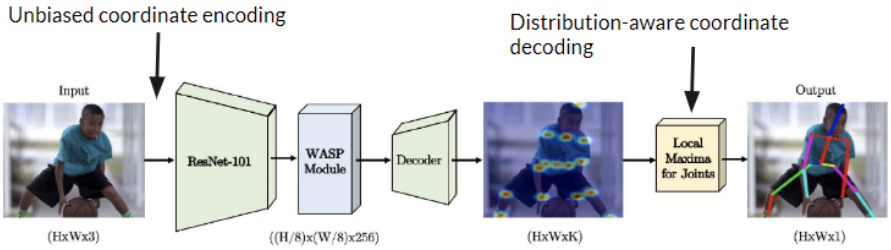

Figure 5: Pipeline of combined UniPose with DARKPose

where λ is stride and $g^{'}$ is the vector of scaled coordinates.

$$G(X; g^{'}) = \frac{1}{2\pi\sigma^2} exp(-\frac{(X - g^{'})^2}{2\sigma^2}) \tag{2}$$

where G is Gaussian.

3.2.2 Decoder module

Decoding is the procedure used to convert the heatmap outputs from the model into keypoints representing each joint. When the heatmap has the same spatial size as the original image, we only need to find the location of the maximal activation as the joint coordinate prediction. But this isn't usually the case as the heatmap outputs are lower in resolution. They have to be upsampled to the image size using a sample specific unconstrained factor λ. This is termed as sub-pixel localisation problem. Generally the pixel is shifted away from the maximal activation towards the second maximal activation by a small margin. While this may work but it is not accurate enough since the heatmaps follow Gaussian distribution. Hence, in DARK the log likelihood of the Gaussian is calculated and optimized with respect to the mean. This will give a more accurate coordinate as it represents the entire distribution of the heatmap. The procedure is summarized in below equations,

$$P(x; \mu, \Sigma) = ln(G) = -ln(2\pi) - \frac{ln(|\Sigma|)}{2} - \frac{(x - \mu)^T \sigma^{-1} (x - \mu)}{2} \tag{3}$$

where P is the log liklehood of the Gaussian G with mean μ and covariance Σ.

$$\mu = m - \Sigma\Sigma^{-1}(m - \mu) \tag{4}$$

After optimizing P using Taylor series the above equation containing μ and m are obtained where m is the maximal activation. Solving this will give accurate coordinates of the keypoints.

3.3 Integrating UniPose with DARKPose

Now we describe our approach to combine Unipose with DARK. DARK representation doesn't really depend on the inherent prediction model rather it only depends on the predicted and groundtruth heatmaps. This property makes it model agnostic. Due to this DARK can be integrated with any model. In our project we tried to integrate with Unipose architecture. During the training the normal ground truth heatmaps are passed through DARK encoder to enhance the keypoint representation before passing them onto the model. During testing the output heatmaps predicted from the model are passed on to DARK decoder in order to get a more accurate representation of the heatmaps. In this way we take the advantage of the DARK representation while keeping the Unipose architecture intact. This is a best of both worlds approach.

4 Experiments and Results

Here we explain the experimental setup of our project.

Figure 6: Samples from the Leeds Sports Dataset.

4.1 Datasets

Leeds Sports Pose Dataset [3]
This dataset contains 2000 pose annotated images of sport persons. The images have been scaled such that the most prominent person is roughly 150 pixels in length. Each image has been annotated with 14 joint locations. Left and right joints are consistently labelled from a person-centric viewpoint. The labels are in a 3x14x2000 matrix with x and y locations and a binary value indicating the visibility of each joint (1 if joint is not visible) The ordering of the joints is as follows:

1. Right ankle
2. Right knee
3. Right hip
4. Left hip
5. Left knee
6. Left ankle
7. Right wrist
8. Right elbow
9. Right shoulder
10. Left shoulder
11. Left elbow
12. Left wrist
13. Neck
14. Head top

The dataset was split into 1000 training images and 1000 validation images. Note: this was according to the paper Unipose. We chose this dataset as it is a very diverse dataset of challenging human poses. At the same time, it is small enough that we can train the model on Datahub. There are also many publicly available benchmarks using this dataset. In further expansions of this work, we can test it on datasets like MPII (25k images), PennAction and BBC Pose (for videos) and COCO Keypoint Dataset.

4.2 Pretrained models comparison

First we compare the performance of two pretrained Unipose models. One with DARK integrated and the other without it. This will serve as an indicator if DARK can enhance the performance just by plugging it during the testing. We use the official weight provided on the Unipose repository finetuned on Leeds sports dataset. We use the validation set from our normal LSP dataset for perfromance

comparison. We use the standard pose estimation metrics like Percentage of Correct Keypoints (PCK) with a threshold of 0.2 of the torso diameter, and the second is PCKh@0.5, which refers to a threshold of 0.5 of the head diameter. We also mean average precision (mAP). The results can be seen in the below table. Here all the metrics show that integrated performance is much higher than normal Unipose model. This can be further improved but we couldn't train more because of our limited computational resources.

Table 1: Pretrained models comparison after 10 epochs

Metric	Without DARK	With DARK
PCK0.2	56.56	63.48
PCKh0.5	54.74	60.82
mAP	38.83	43.24

4.3 Models from scratch comparison

Next we compare the performance of two models trained from scratch. Similar to the above one Unipose model has DARK integration and the other doesn't. We trained the raw Unipose model also from scratch inorder to make a fair assessment of the performance between both the models. We use the train set and validation set taken from the LSP dataset. Similar to the above we use the same metrics. We trained the model for 20 epochs with the hyperparameters shown in Table.2. The results can be seen in table below. We report two sets of results, one is the best results we got across epochs and the other one is the last epoch results. Here for the 1st set of results even though there isn't huge difference, DARK integrated model outperforms it's counterpart. For the 2nd set of results the difference between them increases even more. As discussed above as well this integration shows great promise and given enough computational resources the performance can be increased even more. We also observed that the scores of ankles and knees are almost always the lowest among all 14 keypoints.

Table 2: Hyperparameters

Parameter	value
Weight decay	0.0005
Momentum	0.9
Batch size	8
Learning rate	0.0001
Gamma	0.3333
Step size	13275
Sigma	3
Stride	8

Table 3: Models from scratch comparison best results

Metric	Without DARK	With DARK
PCK0.2	79.53	80.87
PCKh0.5	76.70	76.16
mAP	43.65	45.10

4.4 Ablation study

4.4.1 With only encoder

Now we perform a small ablation study. First we remove the decoder part and use only encoder. For this we train the model for 5 epochs with encoder and validate it without decoder. We note the results obtained.

Table 4: Models from scratch comparison after 20 epochs

Metric	Without DARK	With DARK
PCK0.2	43.68	48.68
PCKh0.5	41.94	45.42
mAP	26.48	26.26

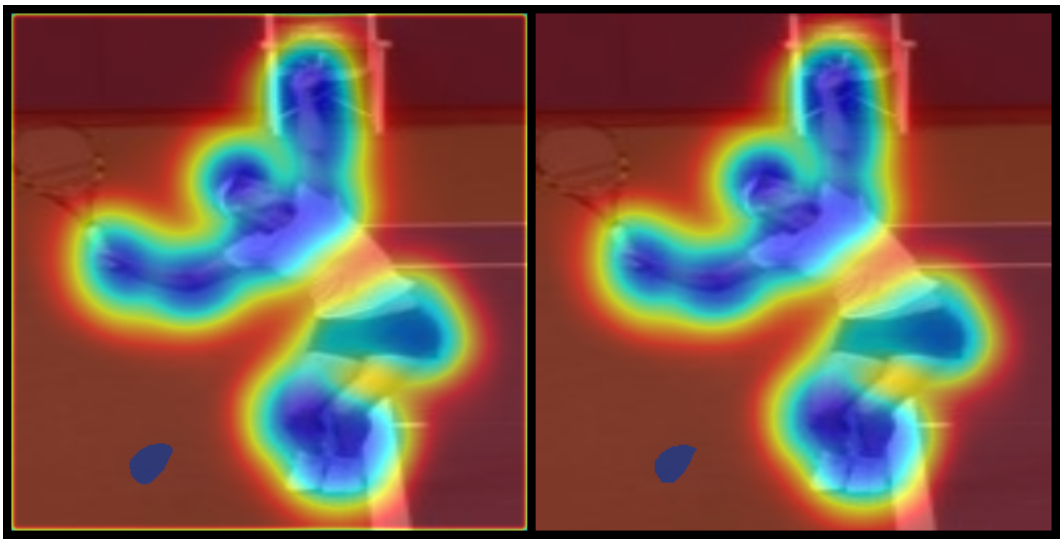

Figure 7: Left: UniPose with DARK Right : UniPose Without DARK. These are the heatmap outputs of the model for an image.

4.4.2 With only decoder

This time we use only the decoder and remove the encoder. We train the model for 5 epochs again without encoder and validate with decoder.

The comparison of both the cases can be seen in the below table. From the results it seems like the effect of decoder is more on the DARK performance compared to encoder. This might also make sense to think that only a minor change is made in the new encoder model when compared to the Unipose encoder model. However the difference is not significant when we keep in mind that this ablation study is performed with only 5 epochs due to computational and time restrictions.

Table 5: Ablation study results

Metric	Without decoder	Without encoder
PCK0.2	73.59	75.28
PCKh0.5	71.76	72.36

5 Future Work

Through this report we showed how adding the DARKPose plugin might improve the performance of UniPose given enought computational resources. In future, we would like to extend the project to include more image datasets like MPII, and video datasets like PennAction and BBCPose. We would also like to implement the improved algorithm, BAPose [11] with DARKPose to do multiple-person pose detection.

6 Supplementary Material

We have attached the code and video demo on gradescope.

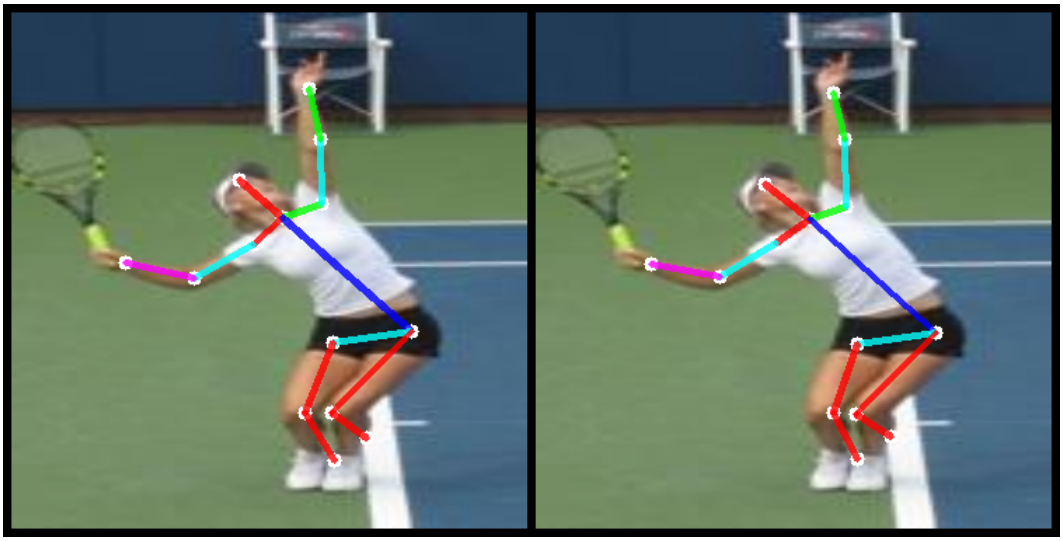

Figure 8: Left: UniPose with DARK Right : UniPose Without DARK. DARK improves the joint localisation by 1-2 pixels. This is evident in the neck area

Figure 9: Left: UniPose with DARK Right : UniPose Without DARK. DARK improves the joint localisation by 1-2 pixels. This is evident in the left elbow area

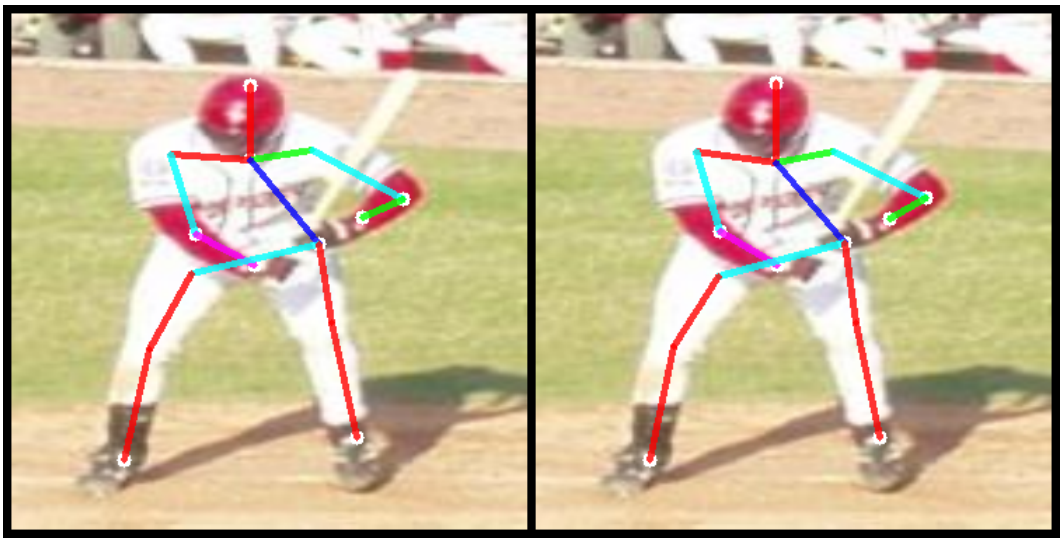

Figure 10: Left: UniPose with DARK Right : UniPose Without DARK. For some joints DARK seems to have more precise coordinates.

Figure 11: Left: UniPose with DARK Right : UniPose Without DARK. For some joints DARK seems to have more precise coordinates.

References

[1] Artacho, Bruno, and Andreas Savakis. "Unipose: Unified human pose estimation in single images and videos." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
[2] Zhang, Feng, et al. "Distribution-aware coordinate representation for human pose estimation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
[3] Sam Johnson and Mark Everingham "Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation" In Proceedings of the 21st British Machine Vision Conference (BMVC2010)
[6] Guanghan Ning and Heng Huang. Lighttrack: A generic framework for online top-down human pose tracking. arXiv preprint arXiv:1905.02822, 2019.
[7] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh. Realtime multiperson 2D pose estimation using part affinity fields. IEEE Computer Vision and Pattern Recognition, 2017.
[8] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation
[9] Gregory Rogez, PhilippeWeinzaepfel, and Cordelia Schmid. LCR-Net: Localization-classification-regression for human pose. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR),
[10] Guanghan Ning and Heng Huang. Lighttrack: A generic framework for online top-down human pose tracking. arXiv preprint arXiv:1905.02822, 2019.
[11] Artacho, Bruno, and Andreas Savakis. "BAPose: Bottom-Up Pose Estimation with Disentangled Waterfall Representations." arXiv preprint arXiv:2112.10716 (2021).
[12] Toshev, A., and Szegedy, C. 2014. Deeppose: Human pose estimation via deep neural networks. In IEEE Conference on Computer Vision and Pattern Recognition [13] Tompson, J. J.; Jain, A.; LeCun, Y.; and Bregler, C. 2014. Joint training of a convolutional network and a graphical model for human pose estimation. In Advances in Neural Information Processing Systems.