

# Super Resolution as a method of improving Human Pose Estimation in Low Resolution Scenarios?

BMVC 2021 Submission # 883

## Abstract

The results obtained from state of the art human pose estimation (HPE) models degrade rapidly when evaluating people of a low resolution, but can super resolution (SR) be used to help mitigate this effect? By using various SR approaches we enhanced two low resolution datasets and evaluated the change in performance of both an object and keypoint detector as well as end-to-end HPE results. We remark the following observations. First we find that for people who were originally depicted at a low resolution (segmentation area in pixels), their keypoint detection performance would improve once SR was applied. Second, the keypoint detection performance gained is dependent on that persons pixel count in the original image prior to any application of SR; keypoint detection performance was improved when SR was applied to people with a small initial segmentation area, but degrades as this becomes larger. To address this we introduced a novel Mask-RCNN approach, utilising a segmentation area threshold to decide when to use SR during the keypoint detection step. This approach achieved the best results on our low resolution datasets for each HPE performance metrics.

## 1 Introduction

Human Pose Estimation (HPE) and keypoint detection are important research topics in computer vision, with many real-world applications such as action recognition and interactive media [19] [15]. Although modern HPE models obtain impressive results on popular datasets such as COCO [18] and MPII [2], their performance degrades substantially when evaluating people of a small scale and low resolution [14]. During keypoint detection, current HPE models utilise Convolutional Neural Networks (CNN). However, as convolutions have a limited robustness to an objects scale [28], ongoing work creating scale invariant CNN architectures remains a key research focus [23] [27] [24] [32]. In contrast, little research exploring how a persons' resolution can be improved for HPE has been undertaken. Super resolution (SR) has been touted within object detection as a panacea for issues of image quality [20] [20] [30], but could it also benefit HPE? This paper will explore multiple SR techniques on low resolution imagery to address this question. By evaluating the performance of a HPE model at different stages of the end-to-end process, we will establish the effect that SR has on HPE and how it varies depending on the target persons' initial resolution.

## 2 Background

### 2.1 Human Pose Estimation

The objective of HPE is to locate and group specific keypoints or bodyparts (shoulder, ankle, etc) from a given image in order to create a human pose [27]. Current HPE methods fall into two categories: bottom-up and top-down approaches. Bottom-up approaches only consist of a keypoint detector, which is used to detect all keypoints in a given image. They then use a grouping algorithm, or human body fitting model, to associate the keypoints with each other in order to create a human pose. By comparison, top-down approaches utilise both an object and keypoint detection component. They start by first detecting the bounding box of each person in an image and then perform keypoint detection inside each bounding box. This negates the need for a grouping algorithm as all the keypoints in each bounding box are assumed to correspond to the same person [7]. As the number of people in a given scene increases, so does the computational cost of top-down approaches. However, this approach is more accurate overall as more people within a scene are detected. **The most popular approach to keypoint detection [8] [22] [27] [27] [9] [60] [8] [8] is known as heatmap regression [8]. A CNN head extracts the initial features and reduces the resolution of the input image, this is followed by a main body which outputs feature maps with the same size as the input feature map, and is followed by a regressor which estimates the predicted heatmaps [27]. Ground truth heatmaps are constructed by using a 2D Gaussian kernel on the given ground truth keypoint and are used to supervise the predicted heatmaps by reducing the L2 loss. Since its introduction by Tompson et al. [24], heatmaps have become the default method for keypoint detection due to its ease of implementation and much higher accuracy than traditional coordinate regression [60] [22].**

### 2.2 Super Resolution

Currently used in multiple real-world applications such as security [69] [25] and medical imaging [13] [12] [54], SR refers to the process of recovering accurate high resolution images from their low resolution counterparts. Modern state of the art SR performance is obtained from deep learning approaches, such as generative adversarial networks (GAN) [53] [16] and auto-encoder architectures [58] [11]. While there exists many different ways of assessing the performance of SR models (structural similarity index [56], feature similarity index [60], etc) the most commonly used metric is the peak signal-to-noise ratio (PSNR). Although PSNR is regularly used as a training metric, the output images generated by maximising PSNR correlates poorly with image quality as perceived by the human eye [9] [55]. This disparity is surprising when findings of recent studies have improved the overall object detection performance in low resolution imagery when combining SR with an object detector [21] [20] [51]. Providing more evidence that deep learning approaches may not perceive image quality the same way as humans, and may in fact learn completely different associations when identifying objects. Some studies however have found a negative impact on object detection performance due to SR if the resolution of the object in the original image is extremely low [28].

046  
047  
048  
049  
050  
051  
052  
053  
054  
055  
056  
057  
058  
059  
060  
061  
062  
063  
064  
065  
066  
067  
068  
069  
070  
071  
072  
073  
074  
075  
076  
077  
078  
079  
080  
081  
082  
083  
084  
085  
086  
087  
088  
089  
090  
\*/091

## 3 Method

While it is difficult to define low resolution with a numerical value, intuition tells us that a low resolution image will be more pixelated and less informative than a high resolution one. We can therefore infer that commonly used computer vision datasets are not low resolution, due to the clarity of the images present. In order to evaluate if SR can improve the HPE results of low resolution people, we used bicubic downsampling to create two low resolution versions ( $\frac{1}{2}$  and  $\frac{1}{4}$  scale) of the COCO validation dataset. We then applied various SR techniques on these low resolution datasets to increase their resolution by a factor of 4. This would then allow us to compare the HPE results between the low resolution images and their SR counterparts. The COCO dataset was chosen for this study as each images annotation also contains the segmentation area (in pixels) of each person in an image. This allowed us to investigate how the effects of SR differ depending on the persons starting segmentation area, as SR may have an adverse effect the lower the initial segmentation area due to the limited amount of starting pixels to reconstruct a high resolution person from. The SR approaches we used to enhance our images were standard bicubic interpolation, ESRGAN [63] and USRNET [68]. For ESRGAN and USRNET we used a GAN version of each model (ESRGAN and USRGAN), and a PSNR maximising version of each model (ESRNET and USRNET). For our HPE model we used HRNET [74], a top-down based approach which achieved one of the highest accuracies across various keypoint datasets at the time of writing. We chose a top-down approach for this study as it consists of both an object and keypoint detection component, which allowed us to test the possible effects that SR has at multiple stages of the end-to-end HPE process.

### 3.1 Object Detection with Super Resolution

The object detector that we used was Faster R-CNN with a resnet-101 and feature pyramid network backbone [75]. This was trained on the standard unaltered version of the COCO training dataset and **the AP and AR performance of this model on the default COCO validation dataset can be seen in Table 1.** As the COCO annotation groups people into a small, medium and large subgroup (S, M and L) depending on that persons given segmentation area, a persons subgroup would usually change when SR is applied. In order to compare the effect that SR has on each subgroup fairly, if someone was defined as small, medium or large in the scaled down image, they would also be defined as that subgroup in the results of the SR image. **It is for this reason we have not reported the S, M, or L results of the default COCO dataset as no fair comparison can be made.** The results showing the average precision (AP) and average recall (AR) of our detector on the low resolution datasets and their SR counterparts can be seen in Table 2 and 3

Dataset	AP	AR
COCO	0.545	0.612

Table 1: The person detection results on the default COCO dataset

Our results show that the overall performance (AP and AR) of the object detector improved once SR was applied. This concurs with previous studies in this area [61] [71] [70]. One recent study however found that the lower the original resolution of the object we are detecting, the worse the object detector would perform after SR was applied [76]. As the small subgroup contained people with a segmentation area between 1 and  $32^2$  pixels, we

Dataset	AP	AP <sub>S</sub>	AP <sub>M</sub>	AP <sub>L</sub>	AR	AR <sub>S</sub>	AR <sub>M</sub>	AR <sub>L</sub>
LR ( $\frac{1}{2}$ scale)	0.507	0.394	0.683	0.752	0.571	0.459	0.741	0.829
Bicubic	0.511	0.399	0.684	0.752	0.577	0.467	0.742	0.828
ESRGAN [83]	0.515	0.402	0.684	0.755	0.581	0.473	0.743	0.829
ESRNET [83]	0.521	0.409	0.692	0.752	0.589	0.483	<b>0.749</b>	0.832
USRGAN [88]	<b>0.522</b>	0.410	0.689	<b>0.756</b>	0.588	0.481	<b>0.749</b>	0.832
USRNET [88]	<b>0.522</b>	<b>0.411</b>	<b>0.690</b>	0.754	<b>0.590</b>	<b>0.485</b>	0.748	<b>0.834</b>

Table 2: The person detection results of the  $\frac{1}{2}$  scale low resolution (LR) dataset and the SR datasets obtained by upscaling the LR dataset by a factor of 4. The best approach for each evaluation metric is highlighted in bold.

Dataset	AP	AP <sub>S</sub>	AP <sub>M</sub>	AP <sub>L</sub>	AR	AR <sub>S</sub>	AR <sub>M</sub>	AR <sub>L</sub>
LR ( $\frac{1}{4}$ scale)	0.387	0.322	0.688	0.716	0.448	0.378	0.762	0.844
Bicubic	0.413	0.351	0.697	0.736	0.478	0.414	0.768	0.868
ESRGAN [83]	0.445	0.385	0.721	0.739	0.509	0.448	0.786	<b>0.891</b>
ESRNET [83]	0.454	0.394	0.728	<b>0.749</b>	0.519	0.459	<b>0.793</b>	0.886
USRGAN [88]	<b>0.456</b>	<b>0.396</b>	<b>0.729</b>	0.743	<b>0.521</b>	<b>0.461</b>	<b>0.793</b>	0.874
USRNET [88]	0.452	0.392	0.725	0.736	0.518	0.458	0.790	0.870

Table 3: The person detection results of the  $\frac{1}{4}$  scale low resolution (LR) dataset and the SR images obtained by upscaling the LR dataset by a factor of 4. The best approach for each evaluation metric is highlighted in bold.

could not confirm whether all people of a smaller segmentation area had improved, simply that this group as a whole did. In order to determine if the improvements in detection rate were skewed by performance variations in subgroups, we conducted a further test. We created 24 new subgroups from our data, grouping people of a similar segmentation area together. The segmentation areas of people within each subgroup (1-24) for our  $\frac{1}{2}$  scale dataset commenced at 1-500 and concluded at 11501-12000 increasing by 500 for each subgroup. For the  $\frac{1}{4}$  scale dataset the segmentation areas of people within each subgroup increased by 125, starting at 1-125 and concluding at 2876-3000. We then evaluated the performance of the object detector across these 24 subgroups. Our findings can be seen in Figure 1 which shows the percentage increase or decrease in object detection rate for each subgroup once SR had been applied.

Contradicting with previous findings, our results demonstrate that the lower the original segmentation area of the object we wish to detect, the more likely it will be detected once SR is performed. Additionally, we found what seems to be a soft threshold of 3000 pixels, as shown by subgroup 6 in the left panel (segmentation area of 2501-3000). People who had an initial segmentation area below this threshold had their detection rate improved once SR was performed on them, aside from bicubic interpolation and ESRGAN for subgroup 2. For people above this threshold however, it is not clear if the overall detection performance would improve or worsen once SR was applied as the results seem to be sporadic in nature. In the right panel the largest subgroup contained people with a segmentation area of between 2876-3000, and as we decrease the subgroup number, and therefore initial segmentation area, there is a gradual improvement in object detection rate.

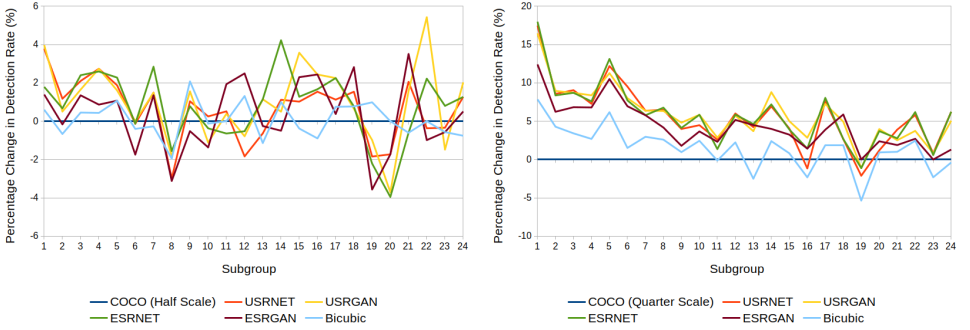


Figure 1: The percentage change in detection rate due to SR for each of our 24 subgroups created from  $\frac{1}{2}$  scale (left panel) and  $\frac{1}{4}$  scale (right panel) datasets and their SR counterparts.

### 3.2 Keypoint Detection with Super Resolution

This section will examine how SR affects the keypoint detection component of the end-to-end HPE process. Our HRNET [7] model was trained on the standard COCO training dataset and the AP and AR performance when evaluating ground truth bounding boxes in the original COCO dataset can be seen in Table 4. As our previous results have shown that the lower the resolution of your object, the better the object detection rate will be once SR had been performed. We now wanted to determine if this also held true for keypoint detection. To eliminate object detection as a variable, we provided HRNET with the ground truth bounding boxes of people in each image. This allowed us to analyse the overall effect that SR has on keypoint detection in low resolution imagery, given that the object detection results are identical. The evaluation metric we used for this study is based on Object Keypoint Similarity (OKS):

$$\frac{\sum_i \exp(-d_i^2/2s^2k_i^2)\delta(v_i > 0)}{\sum_i \delta(v_i > 0)}, \quad (1)$$

where  $d_i$  is the Euclidean distance between the detected and corresponding ground truth keypoint,  $v_i$  is the visibility flag of the ground truth keypoint,  $s$  is the objects scale and  $k_i$  is a per-keypoint constant that controls falloff. In our results we report standard average precision and recall scores [18]: AP, the mean of the AP scores at 10 positions (OKS = 0.50, 0.55, ..., 0.90, 0.95),  $AP_M$  for medium objects,  $AP_L$  for large objects, AR (the mean of AR scores OKS = 0.50, 0.55, ..., 0.90, 0.95) and AR for medium and large people ( $AR_M$  and  $AR_L$  respectively).  $AP_S$  is not reported during keypoint detection as people with a segmentation area  $< 32^2$  do not have their keypoints annotated in the COCO dataset. The results of our keypoint detector on the low and SR datasets can be seen in Tables 5 and 6.

Dataset	AP	AR
COCO	0.765	0.793

Table 4: The AP and AR performance of HRNET [7] on the default COCO dataset when evaluating ground truth bounding boxes.

As our results show, the overall performance of our keypoint detector (AP and AR) improved when evaluating the SR versions of both the  $\frac{1}{2}$  and  $\frac{1}{4}$  scale dataset. When we look

Dataset	AP	AP <sub>M</sub>	AP <sub>L</sub>	AR	AR <sub>M</sub>	AR <sub>L</sub>
COCO $\frac{1}{2}$ Scale	0.722	0.765	<b>0.841</b>	0.752	0.794	<b>0.880</b>
Bicubic	0.728	0.763	0.835	0.760	0.764	0.875
ESRGAN [33]	0.729	0.764	0.825	0.761	0.796	0.866
ESRNET [33]	<b>0.744</b>	<b>0.774</b>	0.831	<b>0.773</b>	<b>0.803</b>	0.873
USRGAN [38]	0.735	0.769	0.826	0.766	0.798	0.870
USRNET [38]	0.741	0.772	0.832	0.772	0.802	0.873

Table 5: The performance of HRNET [27] on the  $\frac{1}{2}$  scale dataset and that same dataset upscaled by a factor of 4 using the various SR techniques. The best result for each evaluation metric is highlighted in bold.

Dataset	AP	AP <sub>M</sub>	AP <sub>L</sub>	AR	AR <sub>M</sub>	AR <sub>L</sub>
COCO $\frac{1}{4}$ Scale	0.538	0.791	0.800	0.573	0.830	0.879
Bicubic	0.601	0.796	<b>0.801</b>	0.637	0.836	0.882
ESRGAN [33]	0.627	0.810	0.786	0.664	0.845	0.875
ESRNET [33]	<b>0.649</b>	<b>0.813</b>	<b>0.801</b>	<b>0.684</b>	0.848	<b>0.888</b>
USRGAN [38]	0.635	0.810	0.794	0.670	0.846	0.882
USRNET [38]	0.647	<b>0.813</b>	0.795	0.681	<b>0.849</b>	<b>0.888</b>

Table 6: The performance of HRNET [27] on the  $\frac{1}{4}$  scale dataset and that same dataset upscaled by a factor of 4 using the various SR techniques. The best result for each evaluation metric is highlighted in bold.

closely however, we can see that simply stating the performance would improve for all observations would be incorrect. Examining the keypoint detection performance for the large subgroup of people (AP<sub>L</sub> and AR<sub>L</sub>) in the  $\frac{1}{2}$  scale dataset, we can see that there was performance degradation as a result of SR. Additionally, not every SR approach we used improved the AP and AR of the medium subgroup of people (AP<sub>M</sub> and AR<sub>M</sub>), as both bicubic interpolation and ESRGAN [33] actually made these results worse. Our findings seemed to hint at a person segmentation area threshold for keypoint detection, above which using SR on a person would worsen the performance of the keypoint detector. To confirm this, we evaluated the change in keypoint detection performance across our 24 subgroups. From this we could determine if there is in-fact an upper limit in object segmentation area, above which the keypoint detection performance would worsen once SR was applied. The results are shown in Figure 2 and 3.

Our figures show that as the initial segmentation area of the people we are evaluating increases, the benefits gained by applying SR for keypoint detection decreases. For our dataset, the threshold beyond which applying SR seems to have a negative affect on keypoint detection, is a segmentation area of between 3501-4000 (subgroup 8 in Figure 2). As beyond this value the percentage change in AP and AR once SR was applied starts to become negative. For people within our smallest subgroup however (subgroup 1 in Figure 3) it is worth noting just how prominent the performance increase is, as we observed a 160% increase in AP performance and 110% increase in AR performance.

3.3 End-to-End Results

This section will evaluate the final keypoint detection performance obtained when using SR for the entire end-to-end top-down HPE process. The results of which can be seen in Table 7

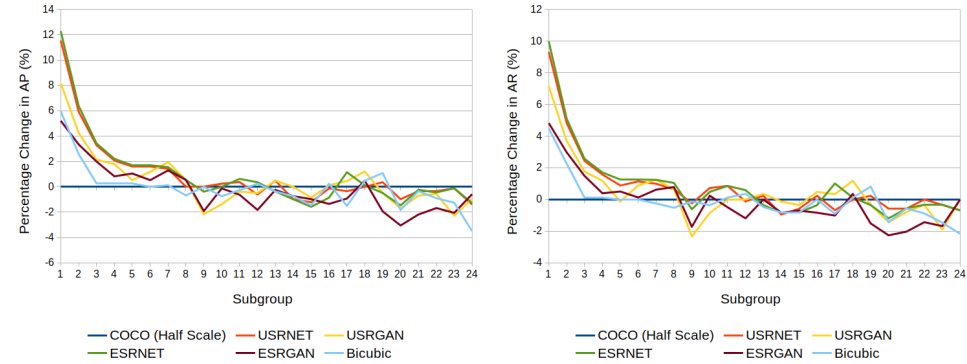


Figure 2: The percentage change in AP (left panel) and AR (right panel) due to SR across our 24 subgroups of the  $\frac{1}{2}$  scale dataset and their SR counterparts.

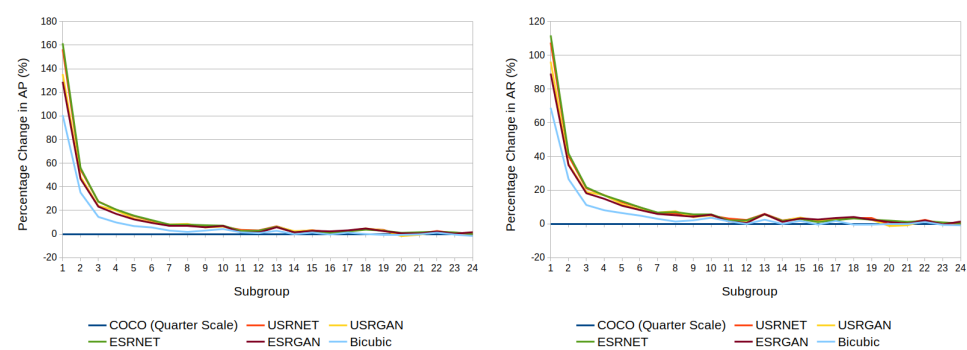


Figure 3: The percentage change in AP (left panel) and AR (right panel) due to SR across our 24 subgroups of the  $\frac{1}{4}$  scale dataset and their SR counterparts.

and 8. Additionally a visualisation showing the increase in HPE performance due to applying SR can be seen in Figure 4

Dataset	AP	AP <sub>M</sub>	AP <sub>L</sub>	AR	AR <sub>M</sub>	AR <sub>L</sub>
COCO $\frac{1}{2}$ Scale	0.704	0.758	<b>0.835</b>	0.747	0.799	<b>0.879</b>
Bicubic	0.709	0.756	0.832	0.753	0.796	0.876
ESRGAN	0.707	0.756	0.832	0.753	0.796	0.876
ESRNET	0.721	<b>0.768</b>	0.827	<b>0.766</b>	<b>0.805</b>	0.872
USRGAN	0.715	0.761	0.827	0.760	0.801	0.874
USRNET	<b>0.722</b>	0.766	0.828	0.766	0.803	0.876

Table 7: The performance of HRNET [27] on the  $\frac{1}{2}$  scale dataset and that same dataset upscaled by a factor of 4 using the various SR techniques. The best result for each evaluation metric is highlighted in bold.

Our results show a clear overall improvement (AP and AR) for keypoint detection when evaluating SR instead of low resolution imagery. What is surprising however is that there is still a performance decrease for the larger people (AP<sub>L</sub>) in our  $\frac{1}{2}$  scale dataset when SR is



Dataset	AP	AP <sub>M</sub>	AP <sub>L</sub>	AR	AR <sub>M</sub>	AR <sub>L</sub>
COCO $\frac{1}{4}$ Scale	0.519	0.785	0.785	0.567	0.833	0.888
Bicubic	0.579	0.791	0.799	0.627	0.836	0.879
ESRGAN	0.602	0.801	0.798	0.649	0.843	0.877
ESRNET	<b>0.630</b>	<b>0.812</b>	0.807	<b>0.676</b>	<b>0.856</b>	0.886
USRGAN	0.613	0.808	0.808	0.661	0.851	0.886
USRNET	0.629	0.811	<b>0.817</b>	0.675	0.852	<b>0.893</b>

Table 8: The performance of HRNET [27] on the  $\frac{1}{4}$  scale dataset and that same dataset upscaled by a factor of 4 using the various SR techniques. The best result for each evaluation metric is highlighted in bold.



(a)  $\frac{1}{2}$  Scale Low Resolution Keypoints



(b) ESRNET SR of (a) Keypoints



(c)  $\frac{1}{4}$  Scale Low Resolution Keypoints



(d) ESRNET SR of (c) Keypoints

Figure 4: A visualisation of the keypoints detected in both a low resolution image and its SR counterpart.

applied. As shown in Table 2, our object detection results for our AP<sub>L</sub> subgroup was either the same or improved once SR was applied. The results in Table 7 however, shows that even though the object detection results have slightly improved, SR has made it harder for our keypoint detector to perform optimally. In other words, it performed worse even with better bounding boxes. As our final contribution for this study we therefore decided to introduce an end-to-end top-down HPE approach that would address this problem.



### 3.4 Mask-RCNN with a Segmentation Area Threshold

By first applying SR to an image, Mask-RCNN [10] would be used to find both the bounding box and segmentation area of people within each image. If the initial segmentation area of a particular person was above a given threshold, then keypoint detection is performed on the original image in a re-scaled bounding box. If the area was below the threshold however, then the SR image is used throughout the end-to-end HPE process. By using one of the best performing SR approaches (USRNET) and the  $\frac{1}{2}$  scale dataset, we performed end-to-end top-down HPE using a segmentation area threshold to decide if the SR or original image should be used during the keypoint detection step. Our threshold chosen was a segmentation area of 3500 or less in the original image, as this is where we began to observe minimal benefits from SR as shown in Figure 2. If the persons segmentation area was below this value then the SR image would be used during keypoint detection. If their segmentation area was greater however, then the original image would be used during keypoint detection instead. The results of our mixed approach when compared to simply using the  $\frac{1}{2}$  scale and USRNET SR dataset can be seen in Table 9.

Dataset	AP	AP <sub>M</sub>	AP <sub>L</sub>	AR	AR <sub>M</sub>	AR <sub>L</sub>
COCO $\frac{1}{2}$ Scale	0.704	0.758	<b>0.835</b>	0.747	0.799	0.879
USRNET	0.722	0.766	0.828	0.766	0.803	0.876
Mixed Approach (W/Threshold)	<b>0.723</b>	<b>0.767</b>	<b>0.835</b>	<b>0.768</b>	<b>0.804</b>	<b>0.882</b>

Table 9: The keypoint detection results from HRNET [11] on the  $\frac{1}{2}$  scale, USRNET and mixed approach datasets.

As the people in the large subgroup all have a segmentation area above the threshold, the AP<sub>L</sub> of the original  $\frac{1}{2}$  scale dataset and mixed approach are now identical. The AR<sub>L</sub> has improved however, and this is due to the our object detector finding more large people in the SR than in the low resolution image, as shown by the increase in AP<sub>L</sub> and AR<sub>L</sub> for USRNET in Table 2. Overall the threshold approach allowed our keypoint detector to perform at its optimum for every evaluation metric, showing that our approach of using Mask-RCNN with a threshold may be a suitable solution for situations where people are both high and low resolution in the same image.

## 4 Conclusion

In this paper we undertook a rigorous empirical study to understand how SR affects the different stages of a top-down HPE process. Prior studies, as well as our initial object detection results, lead us to believe that our final HPE results would also improve once SR was applied; however, this was not the case. Figure 2 shows a clear downward trend, showing that as the initial segmentation area of an object increases, the keypoint detection results after SR decreases. **Additionally, as current state of the art keypoint detectors share a similar feed forward architecture and use the same loss function we see no reason why this observation would differ for a different model, simply the threshold at which the performance decreased would change.** Strangely though, our object detector did not seem to exhibit the same downward pattern; instead the change in object detection rate became sporadic for our larger segmentation area subgroups once SR was applied. **This shows that although both components of a top-down HPE model are reliant on an images resolution to perform optimally, the keypoint detection component relies more on this factor than the object detector,**

whose performance may be more affected by things such as an images context, the lighting of the people in an image, whether people are occluded and if they have a difficult to identify bounding box. Our reasoning for why each components performance degrades as the initial segmentation area increases, is due to the training data that SR models use. Both SR models [58] [63] were trained to reconstruct high resolution imagery from their low resolution counterparts [0], meaning that they have not been trained to reconstruct even higher resolution images from medium resolution counterparts. As we increase the segmentation area of the person we wish to reconstruct, they become increasingly higher in resolution. This causes our SR models to struggle as they has not learnt how to deal with inputs of this size. Although we presented a way to address this problem via our Mask-RCNN approach, the bias introduced by our threshold, as well as not addressing the sporadicity in object detection makes this a sub-optimal solution. Furthermore, our solution only addresses the issue in top-down HPE approaches as it utilises an object detector therefore we would be unable to apply our solution for bottom-up HPE. Future works to find an optimum solution could include an end-to-end HPE model which would learn where in an image to apply SR to, as well as a SR approach which could perform optimally on both low and high resolution objects. Overall however, the improvement in HPE when evaluating the effects of SR on low resolution people is noteworthy, and suggests that SR could be used as a valuable tool for future HPE applications in low resolution scenarios.

## References

- [1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1122–1131, 2017. doi: 10.1109/CVPRW.2017.150.
- [2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.
- [3] Adrian Bulat and Georgios Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 717–732, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46478-7.
- [4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields, 2019.
- [5] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [6] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. pages 7103–7112, 06 2018. doi: 10.1109/CVPR.2018.00742.
- [7] Yucheng Chen, Yingli Tian, and Mingyi He. Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding*, 192:102897, 2020. ISSN 1077-3142. doi: <https://doi.org/10.1016/j.cviu.2019.102897>.

102897. URL <https://www.sciencedirect.com/science/article/pii/S1077314219301778>.
- [8] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [9] Johannes Erfurt, Christian R. Helmrich, Sebastian Bosse, Heiko Schwarz, Detlev Marpe, and Thomas Wiegand. A study of the perceptually weighted peak signal-to-noise ratio (wpsnr) for image compression. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2339–2343, 2019. doi: 10.1109/ICIP.2019.8803307.
- [10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. doi: 10.1109/ICCV.2017.322.
- [11] Xiaodan Hu, Mohamed A. Naiel, Alexander Wong, Mark Lamm, and Paul Fieguth. Runet: A robust unet architecture for image super-resolution. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 505–507, 2019. doi: 10.1109/CVPRW.2019.00073.
- [12] Yawen Huang, Ling Shao, and Alejandro F. Frangi. Simultaneous super-resolution and cross-modality synthesis of 3d medical images using weakly-supervised joint convolutional sparse coding. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5787–5796, 2017. doi: 10.1109/CVPR.2017.613.
- [13] Jithin Saji Isaac and Ramesh Kulkarni. Super resolution techniques for medical image processing. In *2015 International Conference on Technologies for Sustainable Development (ICTSD)*, pages 1–6, 2015. doi: 10.1109/ICTSD.2015.7095900.
- [14] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 196–214, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58545-7.
- [15] N. U. Khan and W. Wan. A review of human pose estimation from single image. In *2018 International Conference on Audio, Language and Image Processing (ICALIP)*, pages 230–236, 2018. doi: 10.1109/ICALIP.2018.8455796.
- [16] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. pages 105–114, 07 2017. doi: 10.1109/CVPR.2017.19.
- [17] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017. doi: 10.1109/CVPR.2017.106.

- [18] 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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.
- [19] D. Luvizon, H. Tabia, and David Picard. Multi-task deep learning for real-time 3d human pose estimation and action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020.
- [20] Bokyoung Na and Geoffrey C Fox. Object detection by a super-resolution method and a convolutional neural networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 2263–2269, 2018. doi: 10.1109/BigData.2018.8622135.
- [21] Bokyun Na and Geoffrey Fox. Object classifications by image super-resolution pre-processing for convolutional neural networks. *Advances in Science, Technology and Engineering Systems Journal*, 5:476–483, 01 2020. doi: 10.25046/aj050261.
- [22] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 483–499, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46484-8.
- [23] Jiquan Ngiam, Zhenghao Chen, Daniel Chia, Pang Koh, Quoc Le, and Andrew Ng. Tiled convolutional neural networks. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL <https://proceedings.neurips.cc/paper/2010/file/01f78be6f7cad02658508fe4616098a9-Paper.pdf>.
- [24] Nanne Noord and Eric Postma. Learning scale-variant and scale-invariant features for deep image classification. *Pattern Recognition*, 61, 02 2016. doi: 10.1016/j.patcog.2016.06.005.
- [25] Pejman Rasti, Tónis Uiboupin, Sergio Escalera, and Gholamreza Anbarjafari. Convolutional neural network super resolution for face recognition in surveillance monitoring. In Francisco José Perales and Josef Kittler, editors, *Articulated Motion and Deformable Objects*, pages 175–184, Cham, 2016. Springer International Publishing. ISBN 978-3-319-41778-3.
- [26] Jacob Shermeyer and Adam Etten. The effects of super-resolution on object detection performance in satellite imagery. pages 1432–1441, 06 2019. doi: 10.1109/CVPRW.2019.00184.
- [27] K. Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5686–5696, 2019.
- [28] Ryo Takahashi, Takashi Matsubara, and Kuniaki Uehara. Scale-invariant recognition by weight-shared cnns in parallel. In Min-Ling Zhang and Yung-Kyun Noh, editors, \*

- Proceedings of the Ninth Asian Conference on Machine Learning, volume 77 of *Proceedings of Machine Learning Research*, pages 295–310. PMLR, 15–17 Nov 2017. URL <http://proceedings.mlr.press/v77/takahashi17a.html>.
- [29] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL <https://proceedings.neurips.cc/paper/2014/file/e744f91c29ec99f0e662c9177946c627-Paper.pdf>.
- [30] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 12 2013. doi: 10.1109/CVPR.2014.214.
- [31] Bin Wang, Tao Lu, and Yanduo Zhang. Feature-driven super-resolution for object detection. In *2020 5th International Conference on Control, Robotics and Cybernetics (CRC)*, pages 211–215, 2020. doi: 10.1109/CRC51253.2020.9253468.
- [32] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. doi: 10.1109/TPAMI.2020.2983686.
- [33] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In Laura Leal-Taixé and Stefan Roth, editors, *Computer Vision – ECCV 2018 Workshops*, pages 63–79, Cham, 2019. Springer International Publishing. ISBN 978-3-030-11021-5.
- [34] Zhihao Wang, Jian Chen, and Steven Hoi. Deep learning for image super-resolution: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP:1–1, 03 2020. doi: 10.1109/TPAMI.2020.2982166.
- [35] Zhou Wang, Alan C. Bovik, and Ligang Lu. Why is image quality assessment so difficult? In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages IV–3313–IV–3316, 2002. doi: 10.1109/ICASSP.2002.5745362.
- [36] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.
- [37] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Learning feature pyramids for human pose estimation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1290–1299, 2017. doi: 10.1109/ICCV.2017.144.
- [38] K. Zhang, L. Gool, and R. Timofte. Deep unfolding network for image super-resolution. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3214–3223, 2020.

[39] Liangpei Zhang, Hongyan Zhang, Huanfeng Shen, and Pingxiang Li. A super-resolution reconstruction algorithm for surveillance images. *Signal Processing*, 90(3):848–859, 2010. ISSN 0165-1684. doi: <https://doi.org/10.1016/j.sigpro.2009.09.002>. URL <https://www.sciencedirect.com/science/article/pii/S0165168409003776>.

[40] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *Image Processing, IEEE Transactions on*, 20:2378 – 2386, 09 2011. doi: 10.1109/TIP.2011.2109730.