HRGAN: Improved Deep High-Resolution Adversarial Learning for Human Pose Estimation

Ting Cao	Siyuan Yan	Gefei Zhang	Luning Li
u7078470	u7050317	u7138112	u7077148

Abstract

In this paper, we improved the performance of the state-of-the-art human pose estimation model High-Resolution Net(HRnet) by solving the implausible human pose estimation problem via our proposed self adversarial learning scheme. We have two main contributions, (i) we leveraged the high-resolution representation rather than encoder-decoder structure, proposed a novel self adversarial training strategy, (ii) we designed a boundary equilibrium scheme for our adversarial training, balanced the learning speed for our discriminator. In our experiment, we empirically proved that our HRGAN outperforms the state-of-the-art top-down model HRNet on the MPII dataset and the COCO keypoint detection dataset. Our network is superior in predicting extreme poses such as small limbs or highly overlapped joints. Our code is available at https://github.com/SkywalkerAtlas/HRGAN.

1 Introduction

Monocular human pose estimation is a challenging task in the computer vision field. The objective is to localize body parts (e.g. ankle, wrist, etc), especially when those body parts are represented in complicated scenes such as severe occlusion or overlapping with other human joints. Many applications can take advantage of human pose estimation, including human action understanding, motion capture, etc. This paper aimed for single-person pose estimation, which is the foundation for multi-person pose estimation [4, 12, 13, 17, 18] and tracking [16, 24].

Many prior deep learning based methods have achieved acceptable performance by regressing the joint heatmap using a top-down approach [1, 7, 15, 22–24]. However, these models struggle with implausible pose estimations under extremely complicated cases. Some researchers tackle the problem by adapting the adversarial training scheme like Generative-Adversarial-Networks(GANs) into their human pose models, estimating poses that are even in highly occlusion circumstances. The advantage of adversarial training is it can give the network a constraint to avoid generating unreasonable poses. [5, 6, 8, 19, 25]. Nevertheless, previous GANs architectures are based on convolutional encoder-decoder structure like Hourglass Network [15], which repeat encode and decode feature information so that learning the rich feature representation. We believe the encoder-decoder based GANs network will lose some useful information to precisely localise human body parts during the process of downsampling and upsampling, causing poor results on small limbs and object edges[6, 8, 19]. Besides, Sun et al. [20] proposed the High-Resolution Net, increased precision by connecting different level resolution subnetworks in parallel thus achieved state-of-the-art results. We adapted this parallel idea into our GAN architecture, tackled the imprecise problem in previous GAN based works as well as produced a more robust result compare to HRNet. That is, we implemented the adversarial training strategy into the HRNet framework to let the discriminator distinguish implausible poses, providing extra information for the generator, thus simultaneously guide it through the training process. We empirically showed our novel joint detection method outperform other methods on two benchmark datasets: the COCO keypoint detection dataset [14] and the MPII Human Pose dataset [1].

Our contributions are:

- We proposed a novel self adversarial learning architecture that leveraged the HRNet, increased estimation accuracy when occlusions and implausible poses are presented.
- We designed a boundary equilibrium scheme for our adversarial training, by balancing the learning speed for our discriminator, we proved that our adversarial training strategy is more stable and can avoid mode collapse when using HRNet as the backbone.

2 Related Work

2.1 Human Pose Estimation

Deep learning based methods are widely applied to the human pose estimation after the introduction of "DeepPose" by [27]. DeepPose [27] used the traditional convolutional structure to regress keypoints of human pose. More studies [26] aims to predict heatmaps which characterize the probability of each keypoints, then followed by choosing the points with the maximum response as the keypoints. Most of the current methods will use heatmaps rather than regress keypoints directly to do the task.

Recent deep learning based methods incorporate more useful tricks with respect to standard convolutional structure. Stacked Hourglass Network [17] is achieving state-of-art performance with the architecture of repeated residual blocks. High-Resolution Net [23] (HRNet) is a more powerful model that connects multi-resolution subnetworks in parallel. HRNet's multi-scale data fusion makes it effective in human pose estimation. Our model uses the design of HRNet as the basis of our network architecture.

2.2 Generative Adversarial Network

Generative Adversarial Network (GAN) is first introduced by Goodfellow et al. [10]. GAN is composed of a generator and a discriminator. They will be trained competitively to get better results compared to training a single network. It comes up with a concern that GAN is hard to train. DCGAN is introduced by [21] and can partially relieve the concern. DCGAN is an all convolutional network, it increases the stability of GAN by eliminating the fully connected layer and adding a batch normalization layer.

DCGAN is not powerful enough to train very deep networks. In order to further solve the stability of GAN, Arjovsky et al. [2] introduced Wasserstein GAN (WGAN), which uses Wasserstein distance as GAN's loss. It further improves the stability of GAN network by solving unreliable gradient problems. BEGAN [3] is developed on the basis of WGAN. It tried to make training GAN easier by balancing between generator and discriminator. It raised an equilibrium formula according to proportional control theory to balance the training period, which is confirmed to have good experimental results on stabilizing GAN's training. We integrate the idea of boundary equilibrium into our training process.

GANs are widely studied in previous works like Nature language Processing, image processing. Discriminators will output a digit ranging from zero to one when applying GANs in those areas to represent the reality of the input. In our human pose estimation task, however, we need the discriminator to judge the reality of heatmaps. Energy-based Gans [30] (EBGANs) treat the discriminator as an energy function rather than zero to one value. We adopt the idea of EBGAN into our model. Our training strategy is largely inherited from Chen et al. [7]. Conditional GAN [16] is introduced to facilitate our discriminator. We also get inspiration from Self Adversarial Training [9] about how they calculate generator discriminator loss and their training settings.

3 Adversarial training with High-Resolution Net

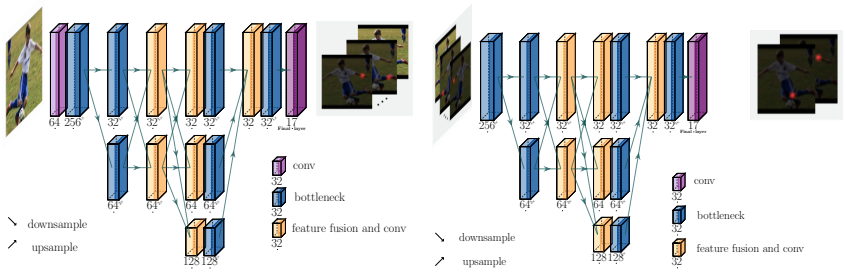
Our HRGAN model has two parts, a generator to produce the human poses and a discriminator to correct wrong poses. Many previous methods like Hourglass Net [15] are repeated top-down and bottom-up structure, which can potentially lose some useful information. To tackle the problem and to do feature fusion better, we use state-of-the-art model High-Resolution Net(HRNet)[20] as the generator. HRNet can keep the high-resolution representation during the whole N-stages network, which performs multi-scale feature fusion repeatedly, receiving different resolution representation information from previous stages and its parallel branches below. After feeding the input images, the generator will output heatmaps that contain confidence scores at each location for different body

parts. For the discriminator, its architecture is very similar to the generator. The difference is that it takes the generated heatmaps along with input images as the input and produces a set of reconstructed heatmaps. By optimizing the reconstructed error, the discriminator can distinguish the real pose from the fake one. The architecture of our generator and discriminator is shown in figure 1 and figure **??**

3.1 Generator

The purpose of the generator is to produce the keypoints heatmaps from its corresponding input image. The generator(HRNet) can learn different scale high-resolution representations and predict precisely keypoints. A content loss and an adversarial loss are added so that the generator can learn spatial feature information and produce more plausible human body poses although in occlusion cases.

High-Resolution Net. We use HRNet as our generator and discriminator. Our HRNet is an N-stage network, the first branch of it is a high-resolution branch. Then in every following stage, an additional new branch with $\frac{1}{2}$ of low resolution is added in parallel to the current branch. In the intermediate stage, the network fuses resolutions from previous stages and different resolutions from its parallel high-to-low branches below. So, the high-resolution branch can keep spatial information and obtain many context and semantic information. Based on this, the predicted heatmaps by HRNet are more accurate and the location for different keypoints is more precise. For human pose tasks, fusing different level feature information is important, the skip-connections[10] is used to connect the low-level features in the shallow layer and high-level features in the deeper layer. Our N-stage HRNet is shown in figure 1.



(a) Generator. (b) Discriminator.

Figure 1: (a) Generator:A HRNet-based network. The bottleneck is a stack of several conv layers of the same size. There are 2 additional subnetworks downsampled to 2× resolution and 4× resolution. The generator produces 17 channel feature maps where each channel is the heatmap of a particular body joint. (b) Discriminator: Similar structure to the generator, the output of the discriminator is seventeen reconstructed heatmaps.

Training strategy for the generator. In the previous work, intermediate supervision is used for training decoder-encoder structure like Hourglass NetNewell et al. [15], it is useful for improving the quality of produced heatmaps and avoid gradient vanishing. While our generator HRNet keeps high-resolution feature representations during the whole network, it can produce high-quality heatmaps without using intermediate supervision, making the network more efficient and easy to train.

The generator is jointly trained with the content loss $\mathcal{L}_{MSE}$ and the adversarial loss $\mathcal{L}_{adv}$ that is back-propagated from the discriminator. The loss of the generator is

$$\mathcal{L}_G = \mathcal{L}_{MSE} + \beta_G \mathcal{L}_{adv}, \tag{1}$$

where β_G is a hyperparameter that is used to decide the weights of $\mathcal{L}_{adv}$.

The final stage of the generator will output N heatmaps that the size of each heatmap is 64x64. Each heatmap contains a confidence score at the location of the jth ground truth keypoint. The MSE loss can make our generator learn useful features that can localize the keypoints. The MSE loss is expressed as

$$\mathcal{L}_{MSE} = \sum_{i=1}^{N} (H_i - \hat{H}_i)^2, \tag{2}$$

where H_i is the ith ground truth heatmap that corresponds to the ith keypoint, and $\hat{H}_i$ is the ith generated heatmap that corresponds to the ith kepoint.

Additional adversarial loss $\mathcal{L}_{adv}$ is also adapted into the generator loss, it is expressed as

$$\mathcal{L}_{adv} = \sum_{i=1}^{N} (\hat{H}_i - D(\hat{H}_i, X))^2, \tag{3}$$

where $\hat{H}_i$ denotes the heatmap generated from the generator, D denotes the discriminator, and X denotes the input image. It pushes the generator to produce more reasonable poses by minimizing the error between the predicted heatmaps and the reconstructed heatmaps. We will discuss more the adversarial loss in section 3.2.

3.2 Discriminator

The prior knowledge of human body structure is important to assign keypoints to the correct location. A human can recognize and correctly assign keypoints although the pose is occluded or is biologically implausible. Some previous works like PIFPAF, adversarial posenet[6, 13] exploit the prior knowledge into their network so that predict human pose better. Inspired by their works, we developed our discriminator to distinguish the real poses from the generated poses. The inputs are either generated heatmaps or ground truth heatmaps, they are concatenated with the input images. The novelty of our discriminator is its outputs are reconstructed heatmaps rather than real or fake. It tries to fit the loss distribution while most GAN tries to match the data distribution. As both low-level features and high-level features are useful to produce a reasonable pose, we use HRNet as the discriminator to learn these features. The discriminator should give suggestions on whether the pose is correct.

Training strategy for the discriminator. We define two objective functions for the discriminator, which are $\mathcal{L}_{real}$ and $\mathcal{L}_{fake}$. If the inputs consist of generated heatmaps, the discriminator will try to reconstruct totally different pose heatmaps, which maximizes the error between reconstructed heatmaps and generated heatmaps. If ground truth heatmaps are contained in the input, the discriminator will try to reconstruct similar ones, which minimizes the error between the reconstructed heatmaps and ground truth heatmaps. The losses we defined are expressed as

$$\begin{aligned} \mathcal{L}_{real} &= \sum_{i=1}^{N} (H_i - D(H_i, X))^2, \\ \mathcal{L}_{fake} &= \sum_{i=1}^{N} (\hat{H}_i - D(\hat{H}_i, X))^2, \\ \mathcal{L}_D &= \mathcal{L}_{real} - r_t \mathcal{L}_{fake}. \end{aligned} \tag{4}$$

The $\mathcal{L}_D$ is the total loss of the discriminator and r_t is used to decide the weight of the adversarial loss. Notice our $\mathcal{L}_{fake}$ in equation 4 is the same as the $\mathcal{L}_{adv}$ in equation 3. The only difference between them is we try to maximize the $\mathcal{L}_{fake}$ while try to minimize the $\mathcal{L}_{adv}$.

For the total loss $\mathcal{L}_D$, it gives each pixel a value, which represents the error between the input and output heatmaps. It has been proved it is helpful to solve the occlusion problem. For example, If one keypoint is very close to the other one, the discriminator will produce a heatmap that one keypoint has a larger error at the location of the other one so that avoiding joint mismatch.

Boundary equilibrium strategy. GAN is hard to train and unstable when the discriminator gets better too quickly. Inspired by previous methods[2, 8], we use a boundary equilibrium strategy. The equation is expressed as

$$k_{t+1} = k_t + \beta_k(\gamma \mathcal{L}_{real} - \mathcal{L}_{fake}), \tag{5}$$

where k_t is bounded between 0 and 1, and β_k is a hyperparameter. We use the term k_t to balance the discriminator and generator. If $\gamma \mathcal{L}_{real}$ is larger than $\mathcal{L}_{fake}$, then the generator will be more powerful than the discriminator. So, k_t will increase, the discriminator will be trained to distinguish fake heatmaps.

Adversarial training. To summary the adversarial training strategy. In figure 2, we minimize the loss of the generator $\mathcal{L}_G$, which consists of the content loss and the adversarial loss. As for the loss of the discriminator $\mathcal{L}_D$, it will maximize the $\mathcal{L}_{fake}$, in the meantime, minimize $\mathcal{L}_{real}$. The discriminator is forced to reconstruct ground-truth heatmaps as its original look and make reconstructed generated heatmaps differs from the generated heatmap. As a result, Large $\mathcal{L}_D$ means unsatisfying generated heatmaps.

The generator and the discriminator are trained in an adversarial way: discriminator tries to maximize $\mathcal{L}_{fake}$ and the generator tries to minimize this term. The network training will be finalized when they come to a compromise. We have another measurement for evaluating the training of our GAN network.

$$\mathcal{M}_{global} = \mathcal{L}(x) + |\gamma\mathcal{L}(x) - \mathcal{L}(G(z_G))| \tag{6}$$

This measurement could serve as a reference to judge the state of GAN model representing whether the network has collapsed or the discriminator loses its discrimination ability.

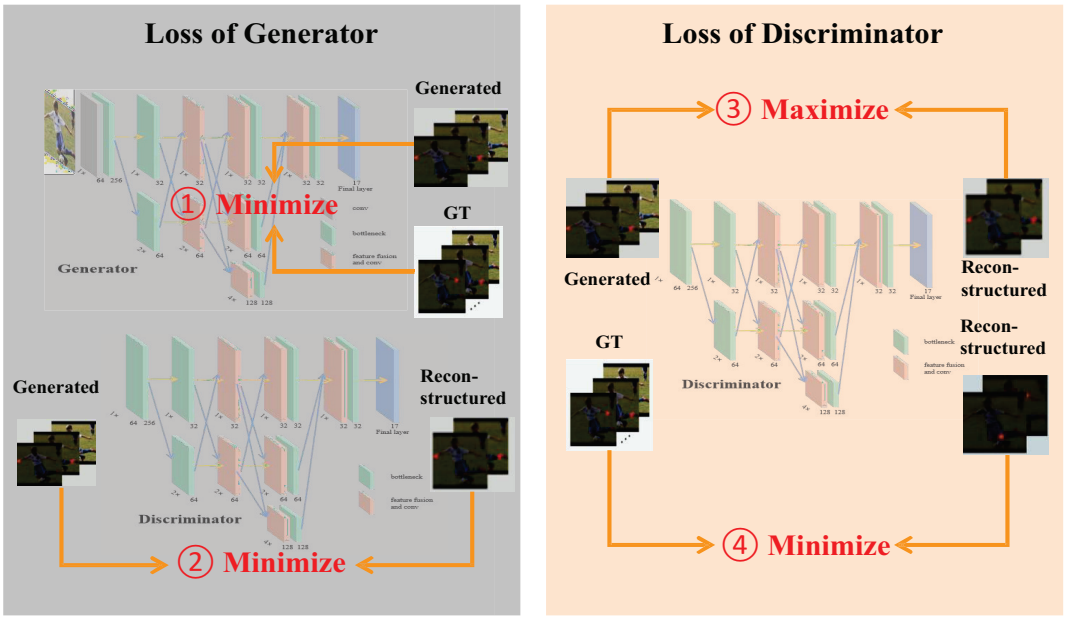


Figure 2: Overview of our Adversarial training strategy: For the generator, it is trained with conventional MSE loss along with adversarial loss. For the discriminator, it maximizes the error between reconstructed heatmap and generated heatmap. Also, it minimizes the error between reconstructed heatmap and ground truth heatmap.

4 Results

We evaluate our model on two benchmark datasets, COCO Dataset [14] and MPII Human Pose Dataset [1]. We follow the preprocessing and data augmentation settings as HRNet [20]. For the optimizer, we take a common approach by using RMSprop [21] and Adam [11] for our generator and discriminator respectively. The learning schedule is similar to the setting in HRNet [20], the initial learning rate is 1e-3, and is dropped by 10 times at the 70th and 120th epochs, the training process is finished at 140 epochs. We trained our model on 4 NVIDIA Tesla V100 with a batch size of 16. Our code is available at https://github.com/SkywalkerAtlas/HRGAN.

4.1 MPII Human Pose Estimation

Dataset. The MPII Human Pose dataset [1] consists of approximately 25K images with 40K annotated poses. For a fair comparison, the input size is resized to 256 $\times$ 256 as other methods.

Evaluation metric. We report the standard PCKh@0.5 (head-normalized probability of correct keypoint) [1] as the metric. A joint is correct if the distance between it and the ground-truth fails within $0.5l$, where l indicates the 60% of the diagonal length of the ground-truth head bounding box.

Testing. For a fair comparison, the test process is identical to HRNet [20], where they followed the standard approach to use the provided ground truth bounding box.

Method	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Total
Chou et al. [8]	98.2	96.8	92.2	88.0	91.3	89.1	84.9	91.8
Newell et al. [15]	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
Cao et al. [5]	98.0	96.8	92.6	88.8	91.4	89.4	**86.7**	92.3
Xiao et al. [24]	98.5	96.6	91.9	87.6	91.1	88.1	84.1	91.5
Sun et al. [20]	**98.6**	96.9	**92.8**	**89.0**	91.5	89.0	85.7	92.3
HRGAN (ours)	97.8	**97.5**	92.5	87.5	**93.9**	**89.7**	85.1	**92.5**

Table 1: Performance comparisons on MPII (PCKh@0.5)

Method	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Total
Bin et al. [3]	**98.9**	**97.6**	**94.6**	**91.2**	93.1	**92.7**	**89.1**	**94.1**
HRGAN (ours)	97.8	97.5	92.5	87.5	**93.9**	89.7	85.1	92.5

Table 2: Performance comparisons with other HRNet based method on MPII (PCKh@0.5)

Result. Table 1 and 2 show the PCKh@0.5 results. Our HRGAN achieves a 92.4 PCKh@0.5 score, outperforms the hourglass network [15] and its GAN variants [5, 6, 8]. Also, our method yield better result compares to its backbone HRNet [20], which shows the effectiveness of our strategy. However, as shown in table 2, compared to the best result reported on MPII [3], our method fails behind, while their method also based on HRNet, their work heavily involve crafted data augmentation process, whereas our method does have any extra data augmentations. We would also like to clarify that our result is produced on the validation set while the compared methods are reported on the test set. The MPII validation set is smaller than the test set, from which our approach could benefit. We can see the visual comparison of HRNet and HRGAN in Figuretab:mpii$_2$.

4.2 COCO Keypoint Detection

Dataset. The COCO Dataset[14] is a comprehensive computer vision dataset with over 200,000 images and 250,000 17 key-points person annotations. Our model has trained on COCO train2017 dataset. It is a large dataset with 57k images and 150k person instances. We resize each human instance bounding box to a fixed size: 256×192 during training. No data augmentation is applied to the dataset. We tested our model on 5000 images COCO val2017 dataset.

Evaluation metric. COCO dataset offered us an official evaluation metric based on Object Keypoint Similarity (OKS). Given the OKS, we can compute AP(Average Precision) and AR(Average Recall). We will record the following metrics as main metrics for our model: AP^{50} , AP^{75}, AP, AP^M, AP^L, and AR.

Results on validation set Our testing is a two-step paradigm: use a person detector to find person instance, use our model to predict keypoints.
We used the same testing scheme as in paper HRNet[20]. Only the output of the generator is considered during inferencing, so our inferencing our model is as efficient as inferencing HRNet.
The result of our method on the COCO validation dataset is listed in Table 3. It is compared with the 8-stage Hourglass model and HRNet-32 models. Our original network outperforms all of the compared networks in all 6 evaluation metrics.
We have an 8.5 AP score improvement compared to Hourglass; 2.0 AP score improvement compared to none-pretrained HRNet; 1.0 AP score improvement compared to pretrained HRNet. The improvement is a credit to our novel model structure and training scheme. We compared visualization results of HRGAN and HRNet in Figure4.

Method	Backbone	Pretrain	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
8-stage Hourglass[15]	8-stage Hourglass	N	66.9	-	-	-	-	-
HRNet-W32	HRNet-W32	N	73.4	89.5	80.7	70.2	80.1	78.9
HRNet-W32	HRNet-W32	Y	74.4	90.5	81.9	70.8	81.0	79.8
HRGAN (ours)	HRNet-W32	Y	**75.4**	**91.0**	**83.4**	**72.3**	**81.6**	**81.0**

Table 3: Performance comparisons on COCO validation set

4.3 Ablation study

We study the influence of each major component in adversarial training on the MPII dataset. All results are produced with identical input settings.

Conventional GAN approach. Conventionally, the output of discriminator is a scalar indicate the confidence of the input being real [9]. Chen et al. [6] proposed an alternative approach for heatmap estimation as they let their discriminator check fidelity for each heatmap channel. We separately implement both strategies, both strategies lead to model collapse, where the generator predicts the same joints for every input image and fails to converge. We believe the model collapse is caused by the few-pixel difference between the real and the fake heatmap in the later training stage, even if they still have differences, the discriminator can not correctly discriminate from the highly compressed features. This result proves the necessity of our adversarial training strategy for measuring reconstructed errors.

Remove boundary equilibrium balance. We remove this balance by setting $\mathcal{L}_D$ back to $(\mathcal{L}_{real} + \mathcal{L}_{fake})/2$. Whereas our original implementation with boundary equilibrium balance can be intuitionally consider as perform WGAN strategy pixel-wisely, this simplified adversarial training can be considered as a degenerated GAN applies on each pixel. We find without boundary equilibrium, the variant can still converge with slower speed, achieves a PCKh@0.5 score of 91.1, which is lower than our original result of 92.5. We believe the reason is boundary equilibrium balance helps to dynamically balance the importance between the generator and the discriminator, thus helps our network to achieve a better result with fewer epochs.

Apply target weight. HRNet [20] has a target weight for each heatmap channel, if the joint is not represent, the corresponding weight will be set to zero. We empirically analyze the effect of integrating it into our discriminator. The framework with target weight reports a 90.8 PCKh@0.5 score. We guess the reason is that our discriminator compares the difference over the entire distributions, thus need all the predicted channels to properly provide enough information for the generator.

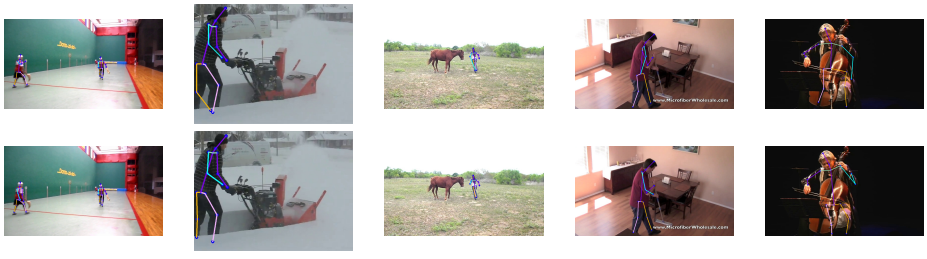


Figure 3: Visual results of HRNet (top) and HRGAN(bottom) on MPII dataset. Our method can corrctly predict joints when occlusion and implausible pose are presented

5 Discussion and Conclusion

In this report, we present an adversarial training scheme for HRNet, reporting more accurate and precise results, especially for the occluded and implausible pose. However, as our framework follows top-down approaches, it still highly rely on the detector, the performance can be negatively affected by highly overlapped or poorly predicted bounding box.

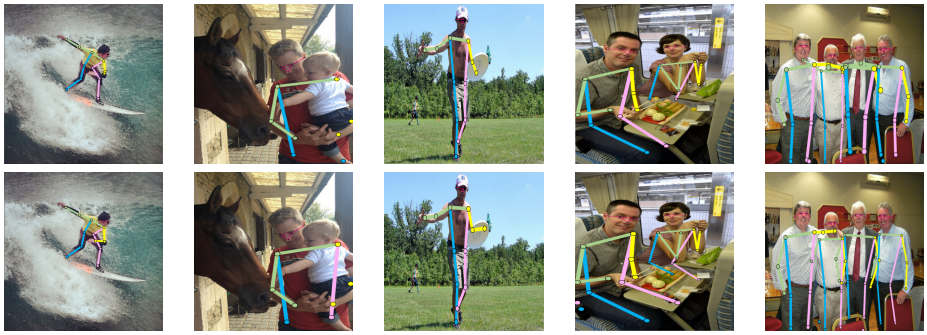


Figure 4: Visualization results on COCO dataset.

Figure 5: Some failure examples of our HRGAN. If the pose is highly overlapped with other people or objects that are similar to joints, our model can still fails.

Since our adversarial training strategy can be integrated into other heatmap-based estimation, one promising future work can be integrating it into bottom-up methods. We believe our promising idea can further increase the performance of human pose estimation.

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[2] D. Berthelot, T. Schumm, and L. Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

[3] Y. Bin, X. Cao, X. Chen, Y. Ge, Y. Tai, C. Wang, J. Li, F. Huang, C. Gao, and N. Sang. Adversarial semantic data augmentation for human pose estimation. *arXiv preprint arXiv:2008.00697*, 2020.

[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[5] Z. Cao, R. Wang, X. Wang, Z. Liu, and X. Zhu. Improving human pose estimation with self-attention generative adversarial networks. In *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 567–572. IEEE, 2019.

[6] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang. Adversarial posenet: A structure-aware convolutional network for human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1212–1221, 2017.

[7] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018.

[8] C.-J. Chou, J.-T. Chien, and H.-T. Chen. Self adversarial training for human pose estimation. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 17–30. IEEE, 2018.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL `http://arxiv.org/abs/1512.03385`.

[11] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. corr abs/1412.6980, 2014.

[12] M. Kocabas, S. Karagoz, and E. Akbas. Multiposenet: Fast multi-person pose estimation using pose residual network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 417–433, 2018.

[13] S. Kreiss, L. Bertoni, and A. Alahi. Pifpaf: Composite fields for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11977–11986, 2019.

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[15] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.

[16] G. Ning, J. Pei, and H. Huang. Lighttrack: A generic framework for online top-down human pose tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1034–1035, 2020.

[17] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy. Towards accurate multi-person pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4903–4911, 2017.

[18] G. Papandreou, T. Zhu, L.-C. Chen, S. Gidaris, J. Tompson, and K. Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 269–286, 2018.

[19] P. Shamsolmoali, M. Zareapoor, H. Zhou, and J. Yang. Amil: Adversarial multi-instance learning for human pose estimation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(1s):1–23, 2020.

[20] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5693–5703, 2019.

[21] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

[22] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient object localization using convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 648–656, 2015.

[23] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

[24] B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018.

[25] A. Zhu, S. Zhang, Y. Huang, F. Hu, R. Cui, and G. Hua. Exploring hard joints mining via hourglass-based generative adversarial network for human pose estimation. *AIP Advances*, 9(3): 035321, 2019.