Supplementary Material for When Human Pose Estimation Meets Robustness: Adversarial Algorithms and Benchmarks

Jiahang Wang [1†] Sheng Jin [2,3] Wentao Liu[4] Weizhong Liu [1] Chen Qian [4] Ping Luo[2]
[1] Huazhong University of Science and Technology [2] The University of Hong Kong
[3] SenseTime Research [4] SenseTime Research and Tetras.AI
jiahangwangchn@gmail.com {jinsheng, liuwentao, qianchen}@sensetime.com
liuweizhong@mail.hust.edu.cn pluo@cs.hku.hk

1. Architecture of Augmentation Generator

We adopt the U-Net [1] architecture to build the augmentation generator, which generates the attention maps for mixing up randomly augmented images. As shown in Figure 1, the augmentation generator is an encoder-decoder with skip connections in between layers, which consists of 6 convolution blocks and 6 transposed convolution blocks. To make sure the size of down-sampled features are equal to up-sampled features for concatenation, we only utilize 5 convolution blocks and 5 transposed convolution blocks for models of input size 384×288 and 128×96.

2. Robustness Enhancement for Different Input Sizes

Taking HRNet-W32 as the backbone network, we conduct more experiments with different input size 128×96, 256×192, and 384×288 on COCO-C to verify the effectiveness of AdvMix for different input resolutions. As shown in Table 1, AdvMix improves both mPC and rPC significantly for all the three different input sizes.

Table 1. **Comparisons** of same backbone with different input sizes between standard training and AdvMix on COCO-C. Results are obtained with the same detection bounding boxes as [2]. We observe that both mPC and rPC are greatly improved, while almost maintaining performance of clean data.

Method	Backbone	Input size	AP*	mPC	rPC
Standard	HRNet-W32	128×96	66.9	47.2	70.6
AdvMix	HRNet-W32	128×96	66.3	48.9	73.8
Standard	HRNet-W32	256×192	74.4	53.0	71.3
AdvMix	HRNet-W32	256×192	74.7	55.5	74.3
Standard	HRNet-W32	384×288	75.7	53.7	70.9
AdvMix	HRNet-W32	384×288	76.2	56.8	74.5

[†]The work was done during an internship at SenseTime Research.

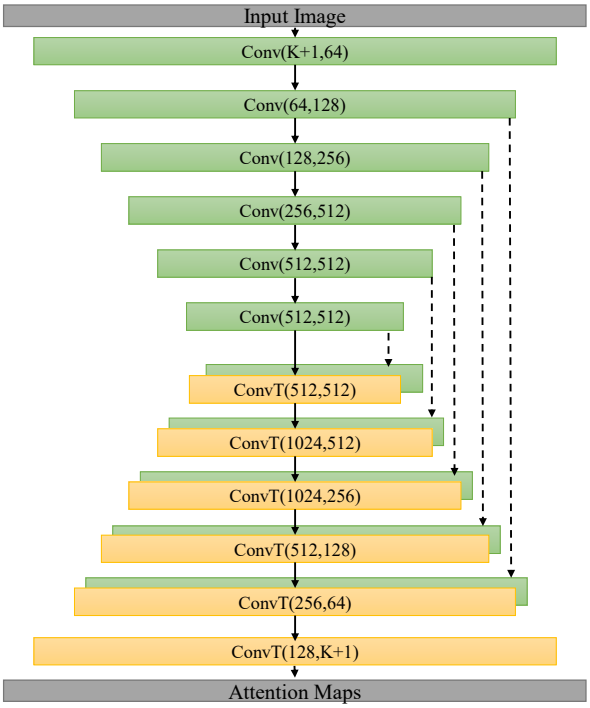

Figure 1. Architecture of Augmentation Generator with input size 256×192. Conv(c,k) means the convolution block with the output channel of c, the kernel size of k. ConvT(c,k) means the transposed convolution block with the output channel of c, the kernel size of k. The kernel size and stride for all blocks are 4 and 2. The activation layer for convolution layers is ReLU, while for transposed convolution layers is LeakyReLU. The black dotted arrow lines between Conv and ConvT denote feature concatenation.

3. Visualization Results

In Figure 2, we provide more human pose results of images with different types of image corruptions, *i.e.* gaussian noise, motion blur, frost and contrast. For each triplet, we

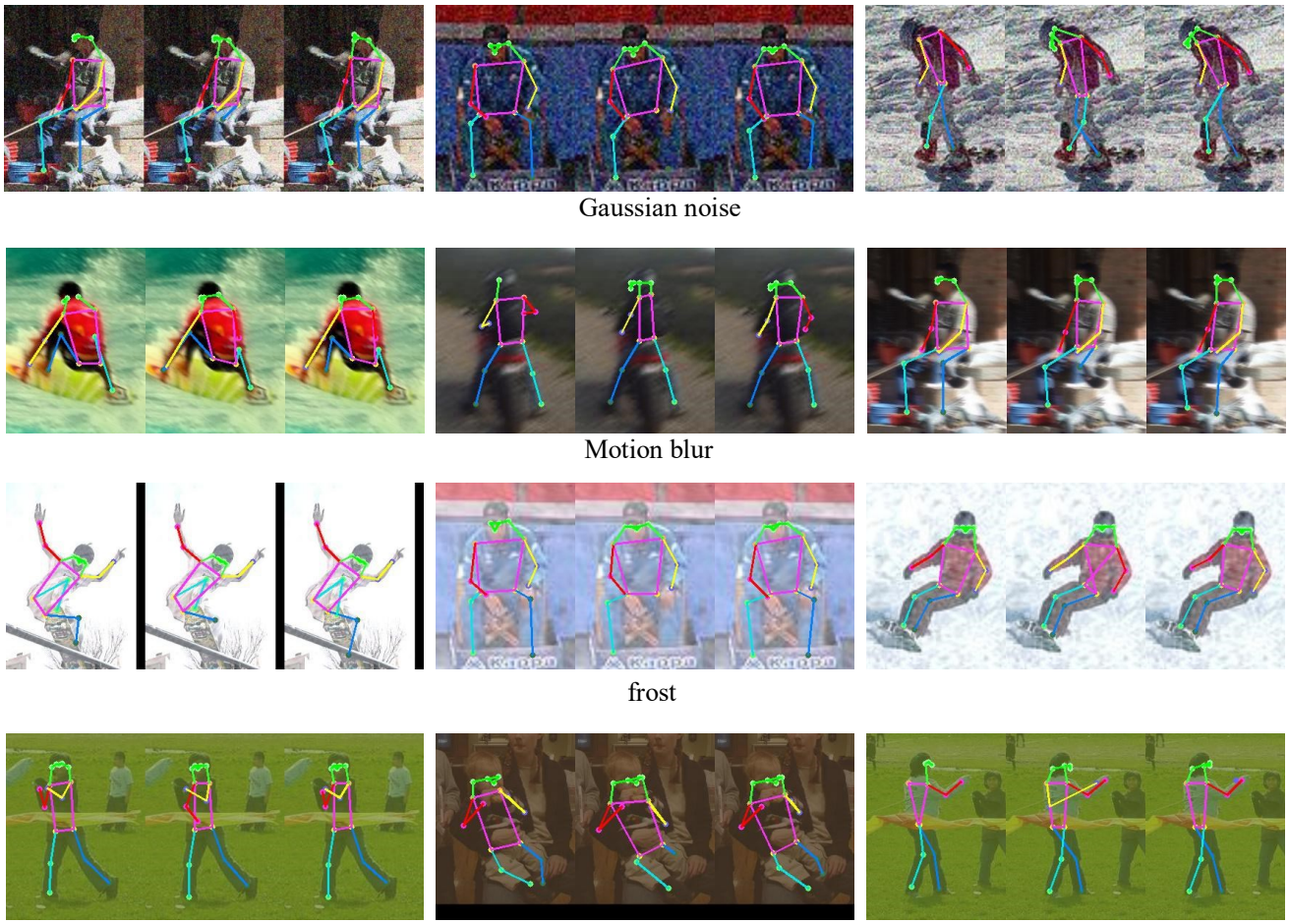

Figure 2. Qualitative comparison between HRNet without and with AdvMix. For each image triplet, the images from left to right are ground truth, predicted results of Standard HRNet-W32, and predicted results of HRNet-W32 with AdvMix.

visualize the ground-truth (the left column), the prediction of the Standard HRNet-W32 (the middle column), and the prediction of HRNet with AdvMix (the right column). We observe that 1) the standard pose models suffer large performance drop on corrupted data, and 2) models trained with AdvMix perform consistently better than the baseline methods on various corruptions.

References

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1

[2] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. *arXiv preprint arXiv:1902.09212*, 2019. 1