Adversarial Synthesis of Human Pose From Text

Yifei Zhang[1,2], Rania Briq[1], Julian Tanke[1], and Juergen Gall[1]

[1] Computer Vision Group, University of Bonn
briq,tanke,gall@iai.uni-bonn.de
[2] Bonn-Aachen International Center for Information Technology, RWTH-Aachen University
yifei.zhang@rwth-aachen.de

1 Supplementary

Loss Curves. In WGANs, the loss is known to be an indication of the quality of the generated samples where its value indicates the distance to the true distribution. In Fig. 2, we show the loss curves of the generators of WGAN-GP and WGAN-LP. We can observe that the loss of two WGAN models decreases (in the absolute value) across the training iterations, indicating that the generator is learning to generate plausible poses and is improving over time. However, we observe that the loss curve of WGAN-GP decreases slightly less than WGAN-LP and more slowly, especially in the second training phase where $\lambda = 150$ compared to 10 in the first phase, and such a large value has been shown to deteriorate training substantially [19], although in our results the deterioration is not substantial.

Qualitative Results. We also include additional qualitative results to point out the differences in the synthesized poses stemming from changing the underlying GAN model. Fig. 3 shows the synthesized poses of Vanilla GAN. While sometimes the poses look realistic and consistent with the input text, changing the noise vector resulted often in very unrealistic poses due to mode collapse. In both WGAN variants, the results look much better than the vanilla GAN.
Fig. 4 corresponds to the WGAN-GP.

Subject Pose. We are also interested in how the generated pose changes based on the subject, e.g. an adult versus a young person. Fig. 5 shows some generated poses for the same caption with different subjects while keeping the noise fixed. The generated poses show subtle difference between the different genders. But if we look more closely, we can find that a young person's pose is slightly smaller than an adult's pose, which reflects the reality. This indicates that for pose generation, the subject of the caption does not matter very much, and what really matters is the action.

Quantitative Results. In Fig. 6, 8 and 7, we plot the pose distance histograms corresponding to table 1,2 from the manuscript for the WGAN-LP, regression and vanilla GAN and WGAN-LP to show the distribution of distances. In Fig. 6,

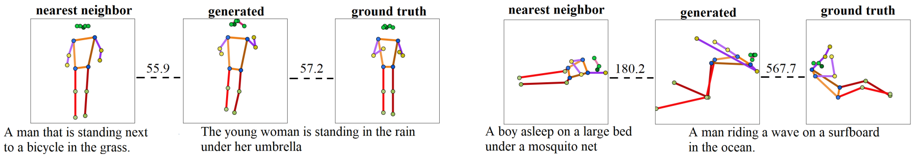
nearest neighbor
generated
ground truth
55.9
57.2
A man that is standing next to a bicycle in the grass.
The young woman is standing in the rain under her umbrella
nearest neighbor
generated
ground truth
180.2
567.7
A boy asleep on a large bed under a mosquito net
A man riding a wave on a surfboard in the ocean.

Fig. 1. Two generated poses, their ground truth poses, and their nearest neighbor poses in the validation set. The text descriptions are below the poses and the distances are shown between them. Left: the ground truth is close to the generated pose and the nearest neighbor has a similar text description. Right: the ground truth is far from the generated pose and the nearest neighbor has a very different text description. However, the large distance to the ground truth is due to the opposite orientation of the pose.

we can see that the generated poses' distances to their NN poses (blue and orange) are much smaller than their distances to all poses on average (purple and brown), while their distances to their ground truth (green) and text-NN (red) poses are shifted away from the average distances towards the NN distances, meaning that in the model the text encodings are indeed guiding the poses synthesis toward the correct direction. For the regression (Fig. 8), such phenomenon is less evident. For the Vanilla GAN (Fig. 7), such phenomenon is even much less evident.

In Fig. 1, we show why a generated pose is sometimes far from the ground truth pose, even though it looks plausible for the given input text.

Noise Interpolation Test. Similarly to the text interpolation test, we also perform a noise interpolation test, where the text is kept fixed and the noise vector is interpolated. As in the text interpolation test, we observe smooth transitions over the interpolated noise vector in Fig. 9.

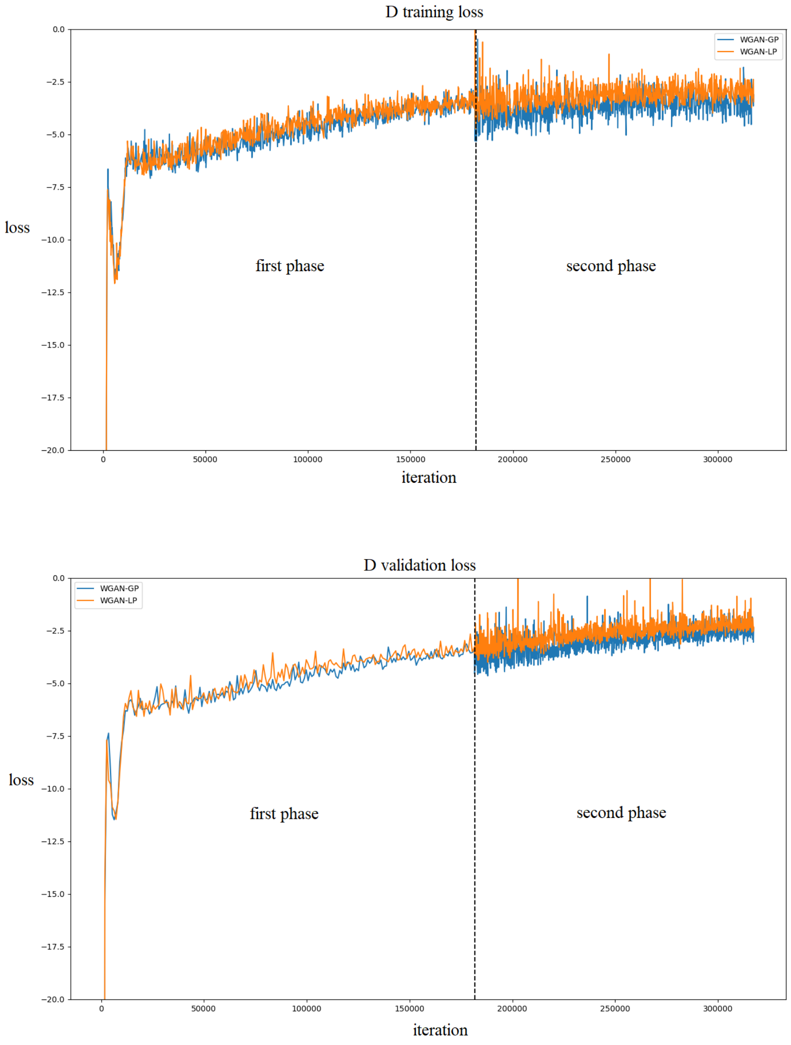

Fig. 2. The training and validation loss curves of the critic D during the two training phases for the two WGAN variants. The orange and blue curve correspond to WGAN-LP and WGAN-GP, respectively.

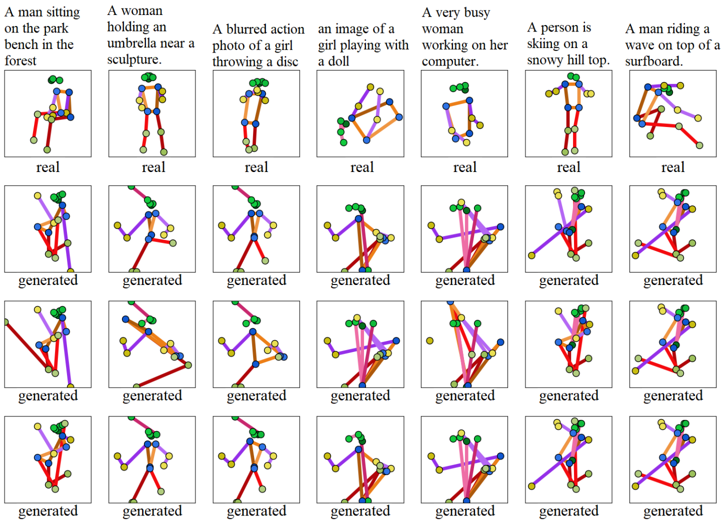
A man sitting on the park bench in the forest
A woman holding an umbrella near a sculpture.
A blurred action photo of a girl throwing a disc
an image of a girl playing with a doll
A very busy woman working on her computer.
A person is skiing on a snowy hill top.
A man riding a wave on top of a surfboard.
real
generated

Fig. 3. Some sample outputs of the model trained with the Vanilla GAN. The first row is the ground-truth from the validation set. The text on the top is the associated text. The three poses below each real pose are synthesized by the model from the text on the top with different noise vectors z.

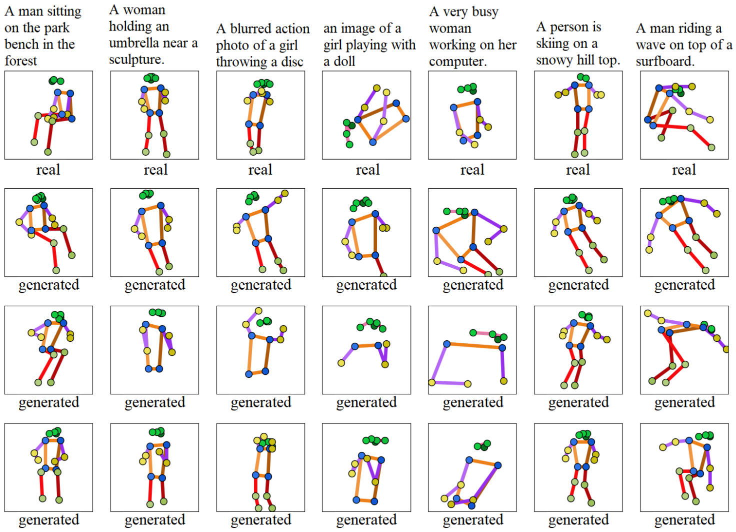
A man sitting on the park bench in the forest
A woman holding an umbrella near a sculpture.
A blurred action photo of a girl throwing a disc
an image of a girl playing with a doll
A very busy woman working on her computer.
A person is skiing on a snowy hill top.
A man riding a wave on top of a surfboard.
real
generated

Fig. 4. Some sample outputs of the model trained with the GP term (WGAN-GP).

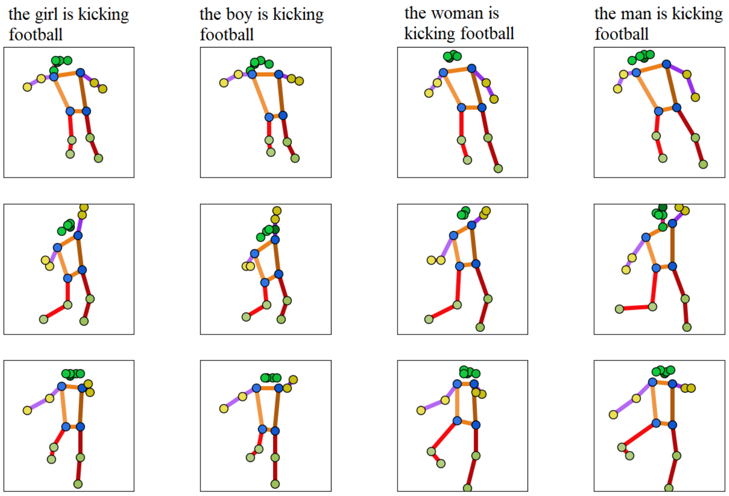
the girl is kicking football
the boy is kicking football
the woman is kicking football
the man is kicking football

Fig. 5. Poses synthesized from captions with different subject genders and age. The caption to synthesize each column of poses is on the top. The noise input is the same for each row.

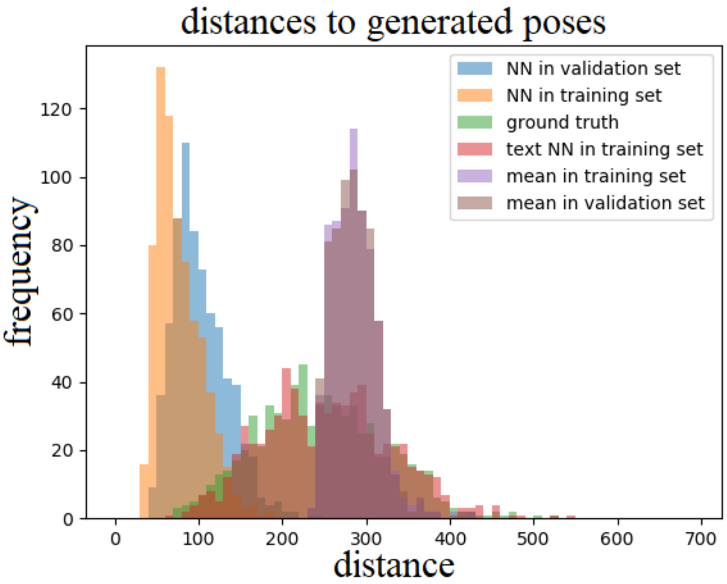

Fig. 6. WGAN-LP. Histograms of pose distances.

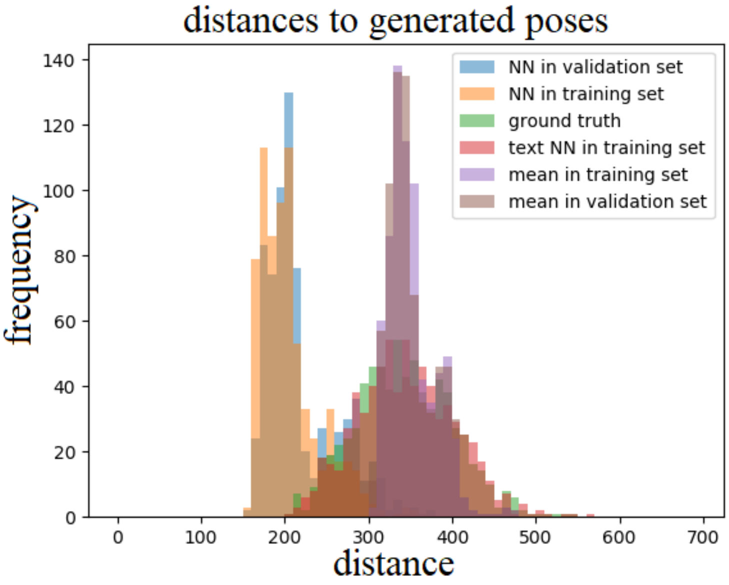

Fig. 7. Vanilla GAN. Histograms of pose distances.

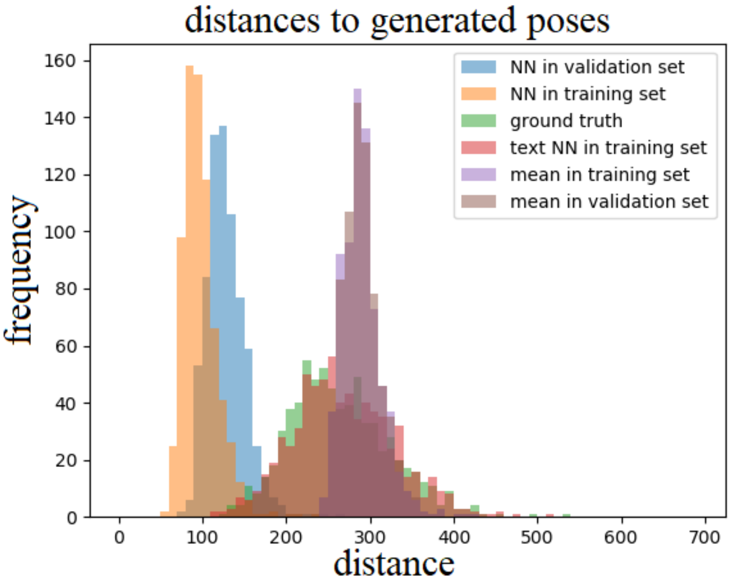

Fig. 8. WGAN-LP Regression. Histograms of pose distances.

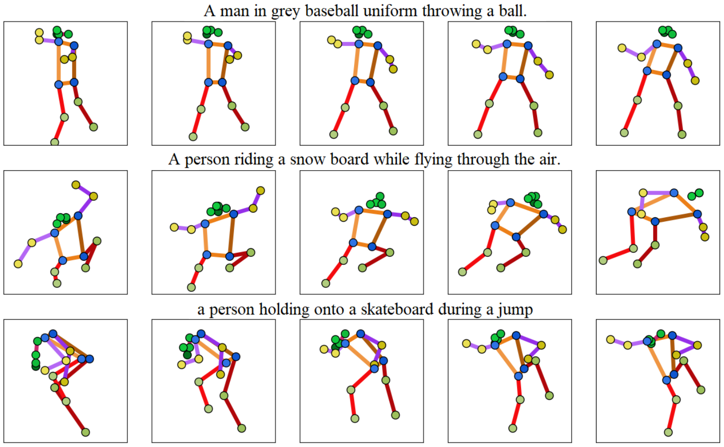
A man in grey baseball uniform throwing a ball.
A person riding a snow board while flying through the air.
a person holding onto a skateboard during a jump

Fig. 9. Interpolation results of noise input. In each row, the five poses are synthesized from the text on the top. The noise inputs of the three poses in the middle are interpolated between the noise inputs of the leftmost and rightmost poses.