Supplemental Material: Hallucinating Pose-Compatible Scenes

Tim Brooks, Alexei A. Efros

UC Berkeley

A Humans in Context Meta-dataset Details

Our dataset contains diverse footage of humans immersed in everyday environments. Each image is supplemented with pseudo-ground truth human pose attained using OpenPose [4,3]. The data is sourced from 10 existing human and action recognition datasets, with the numbers of clips and frames from each source dataset detailed in Table 1. Video footage provides a vast source of diverse human activity, and ensures all poses are represented, rather than only human poses photographers choose to capture in still images. For the MPII [2] dataset, which is primarily a still image dataset, we use short video clips of the frames preceding and following each image.

Each dataset contains unique biases, and combining data sources is less subject to the bias of any particular dataset. Different datasets also offer different types of scenes. For example, Moments [20] includes classes absent from HVU [6], and Oops [8] contains uncommon accidental actions. The number of useful examples from each source was only evident after extensive curation.

We filter out videos where either dimension is shorter than 256 pixels, and we resize remaining videos using Lanczos resampling [7] such that the smaller edge is exactly 256 pixels. We exclude videos with an average bitrate below 0.9 bits per pixel, or with a framerate that does not fall between (and cannot be subsempled to fall between) 23.9 fps and 30 fps. Videos are truncated to 3000 frames. Source datasets which provide pre-extracted frames only undergo quality filtering by spatial resolution.

Frames are then filtered to contain a single person using pretrained Keypoint R-CNN [11,23] person detection. Person bounding boxes are detected for each frame, with a minimum accuracy of 95%, a minimum bounding box area of 1% of the total frame area, and non-maximum suppression of overlapping bounding boxes with an intersection over union greater than 0.3. With these thresholds, any frame with more than a single person detected is removed. Stricter thresholds are then applied to the remaining frames with a single person bounding box: a minimum accuracy of 98%, a minimum bounding box area of 4% of the total frame area, and a maximum bounding box area of 80% of the total frame area. These thresholds ensure with high accuracy that there is a single person present in the frame at a reasonable size. Frames are then cropped to a 256×256 resolution toward the average bounding box center for each contiguous segment of frames.

Table 1: **Humans in Context source data.** Our dataset consists of video clips filtered from 10 existing human and action recognition datasets. High quality clips have sufficient bitrate, framerate and resolution. Person clips are those where pretrained person detection and pose prediction networks assert that a single person is present. In total we curate 229,595 clips and 19,503,700 frames.

	# Video Clips			# Frames	
	Source	High Quality	Person	High Quality	Person
HVU [6]	566,489	353,174	105,634	98,603,223	9,590,407
Moments [20]	757,804	653,368	54,156	56,074,418	3,374,112
Kinetics-700-2020 [17,5]	620,119	432,502	26,911	78,037,500	2,428,079
Charades [22]	9,848	7,319	16,967	6,256,421	2,157,074
InstaVariety [13]	2,545	2,449	5,773	1,898,824	730,211
Oops [8]	29,940	27,953	8,360	5,738,042	596,488
MPII [2]	24,987	24,980	8,820	1,025,459	352,498
VLOG-people [9]	663	555	1,261	321,071	163,956
PennAction [26]	2,326	2,221	1,208	161,029	76,112
YouTube-VOS [24]	4,519	4,511	505	613,441	34,763
Total	2,019,240	1,509,032	**229,595**	248,729,428	**19,503,700**

Pseudo-ground truth pose labels are computed for each frame using OpenPose [4,3] keypoint prediction. We use the single-scale OpenPose version to compute 18 body keypoints. Similar to person detection, we use a relaxed total score threshold of 2.5 when filtering for multiple people, and a strict total score threshold of 10.0 when ensuring there is a single person. Each individual keypoint has a score threshold of 0.3, and keypoints below this threshold are marked as not visible in the frame. To avoid frames of just legs or torso, we only include frames where the keypoint at the base of the neck is visible, and where a total of at least 8 of 14 keypoints (excluding eyes and ears) are visible.

The final dataset only includes clips of at least 30 adjacent frames where each frame passed filtering. Note that multiple clips may be sourced from the same video, and that duplicate videos from different source datasets are possible.

A.1 Dataset licenses

The HVU dataset [6] is released for non-commercial research and educational purposes only, and was attained directly from the dataset authors. The Moments in Time [20] dataset is released for non-commercial research and educational purposes, and was attained from the dataset project website. The Kinetics dataset [17,5] is licensed by Google Inc. under a Creative Commons Attribution 4.0 International License, and videos were downloaded directly from YouTube. The Charades dataset [22] is released under a non-commercial license detailed here: https://prior.allenai.org/projects/data/charades/license.txt; data was downloaded from the project webpage. The InstaVariety dataset [13] is released for non-commercial academic use, and was attained

directly from the dataset authors. The Oops dataset [8] is released under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, and was downloaded from the project webpage. The MPII [2] dataset is released under the Simplified BSD License detailed here: https://github.com/peiyunh/rg-mpii/blob/master/data/mpii_human/annotation/bsd.txt; data was downloaded from the project webpage. The VLOG dataset [9] is released for non-commercial research purposes only, and was downloaded from the project webpage. The PennAction dataset [26] is released without a license and was downloaded from the project webpage; dataset authors confirmed there are no terms of use and only ask the corresponding paper [26] be cited. The YouTube-VOS dataset [24] is released for non-commercial research purposes only and was downloaded from the project challenge webpage.

B Model Implementation Details

We train all models at 128×128 resolution. Many aspects of our model are borrowed directly from StyleGAN2 [16], including non-saturating logistic loss [10], equalized learning rates for all parameters [14], R_1 regularization [19], path length regularization [16], and exponential moving average of generator parameters [14].

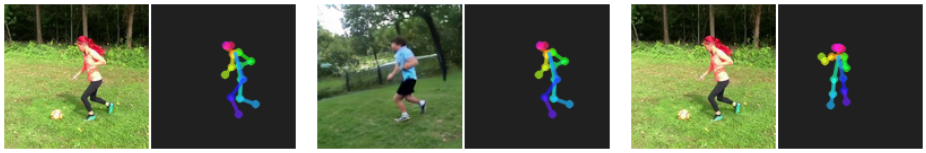
We use a learning rate of 2.5×10^{-3}, an exponential moving average rate of $\beta = 0.995$, a moving average warmup of 150,000 steps, and R_1 regularization strength of $\gamma = 0.05$. We remove spatial noise maps to isolate control over the scene to the latent code. We also remove style mixing regularization during training. Our final model was trained with a minibatch size of 120 on 10× NVIDIA Quadro RTX GPUs, and for 1,000,000 steps. Our ablations trained for 1 week, and we let the final large model continue for 3 weeks. The large generator has 85.4M parameters and discriminator has 98.2M. Ablations and the Pix2PixHD baseline were trained with a batch size of 40 on 5× NVIDIA GeForce RTX 2080 GPUs for 600,000 iterations. Multiple checkpoints were saved throughout training, and the checkpoint with the lowest FID score was used for all evaluation. The Pix2Pix baseline was trained for 10,000,000 iterations with a batch size of 40 on 5× NVIDIA GeForce RTX 2080 GPUs .

Code used from the public implementation of StyleGAN2-Ada is released under the NVIDIA code license found here: https://github.com/NVlabs/stylegan2-ada/blob/main/LICENSE.txt. Code used to run the Pix2Pix baseline is released under the BSD License license found here: https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix/blob/master/LICENSE.

B.1 Data augmentation

Data augmentation of both generated and real images just prior to the discriminator can improve robustness and prevent the discriminator from overfitting to the train dataset [15,27]. Our augmentation parameters are largely based on [27]. Brightness is augmented by randomly offsetting intensity by a value uniformly

Fig. 1: **Data augmentation.** We apply random spatial, cutout and color augmentations to frames and poses just prior to the discriminator network. Each pair above shows the original frame and pose on the left and the augmented output on the right.



(a) Real (b) Generated (c) Mismatched

Fig. 2: **Mismatch discrimination.** The discriminator must classify (a) a real image and pose as real, (b) a generated frame and conditional pose as fake, and (c) a real frame and mismatched pose as fake.

sampled from -25% to $+25\%$. Saturation is augmented by interpolating red, green and blue channels toward or away from the mean of all three at each pixel, with interpolation weights uniformly sampled from 0.0 to 2.0. Contrast is augmented by interpolating color values toward or away from the mean of all color values in an entire frame sequence, with interpolation weights uniformly sampled from 0.5 to 1.5. Horizontal flipping is applied with a 50% chance to all frames and poses in a sequence. Frames are scaled by a factor uniformly sampled from 0.8 to 1.25 and translated by an offset sampled uniformly from -12.5% to $+12.5\%$. A random cutout, half the size of each dimension and randomly placed, is erased from each frame. Spatial transformations applied to frames are also applied to poses so that the frames and poses still correspond correctly. We briefly experimented with dropout augmentation of pose, but did not find it helpful. See Figure 1 for examples of our data augmentation.

B.2 Mismatch discrimination

We force the discriminator to pay attention to pose conditioning by providing a mismatched real image with the incorrect pose conditioning as an additional fake example. For the mismatched fake example, the pose embedding and keypoint heatmaps both take pose from another sample in the minibatch. This training method was first introduced in text-to-image generation [21] but has not been widely used in the image or video translation literature; we found training with mismatch discrimination provides a slight improvement, forcing the discriminator to use conditioning.

B.3 Human disentanglement

Our generator can be used to place a human subject in a new scene, as we outline in Section 5.3 of the main paper. We accomplish this by optimizing for a latent code which produces a scene matching one image and a subject matching another. We separately optimize the latent code used at each *scale* of our model, which is similar to the $\mathcal{W}^+$ space [1] used for inversion (although slightly lower dimensional). We minimize perceptual loss [12,25] between a subject-only crop of the first generated image and the composition. When generating subject-only images, we zero out the learned constant input to the StyleGAN2 generator [16], which we found helps isolate the subject from the background. The crop region is attained from human pose. We also minimize perceptual loss between scene-only versions of the second generated image and composition image. We optimize for 1000 steps using the Adam optimizer [18] and a learning rate of 0.05.

C Additional Results

Please see Figures 3 4 5 6 7 for random uncurated samples from our model.

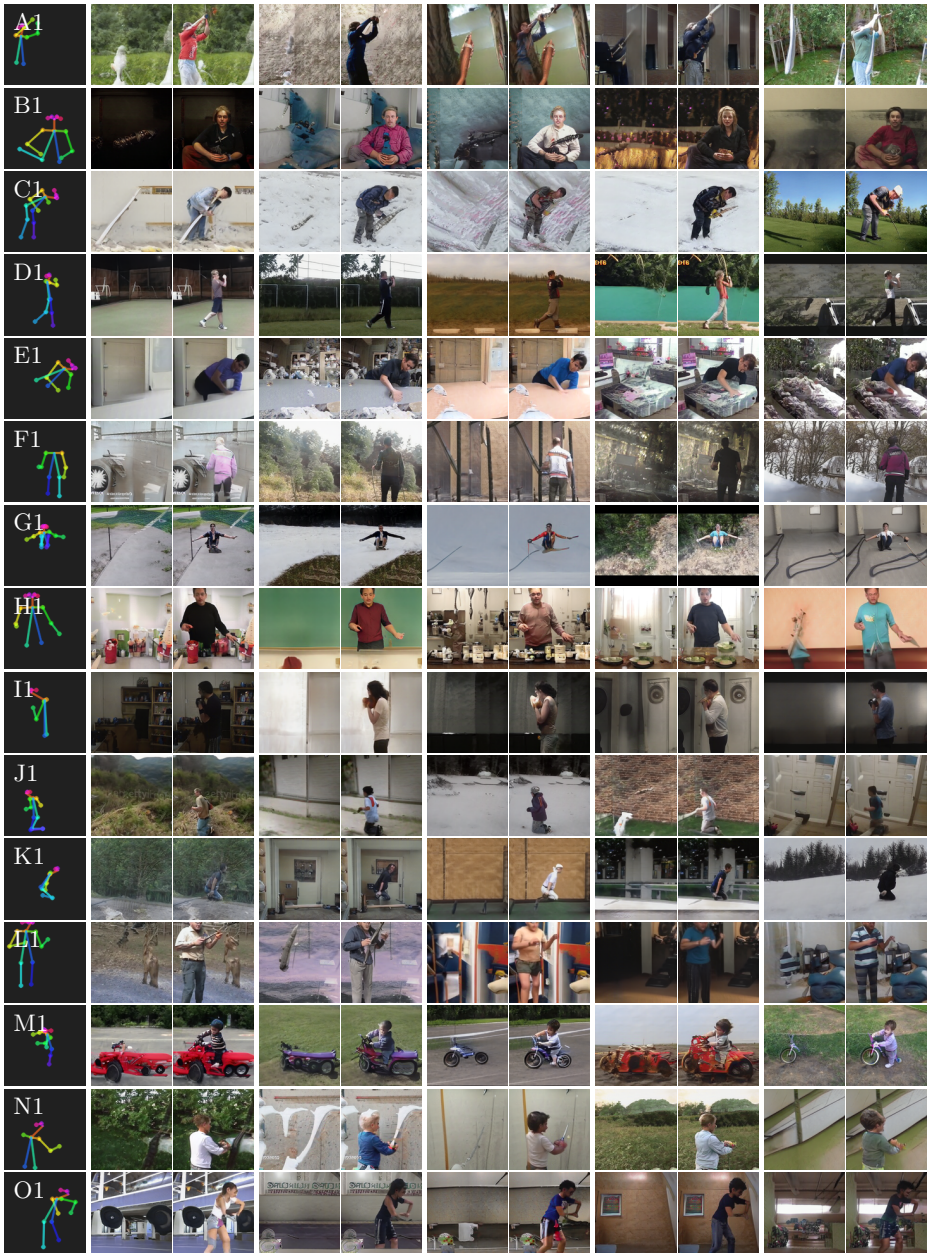


Fig. 3: **Random samples.**

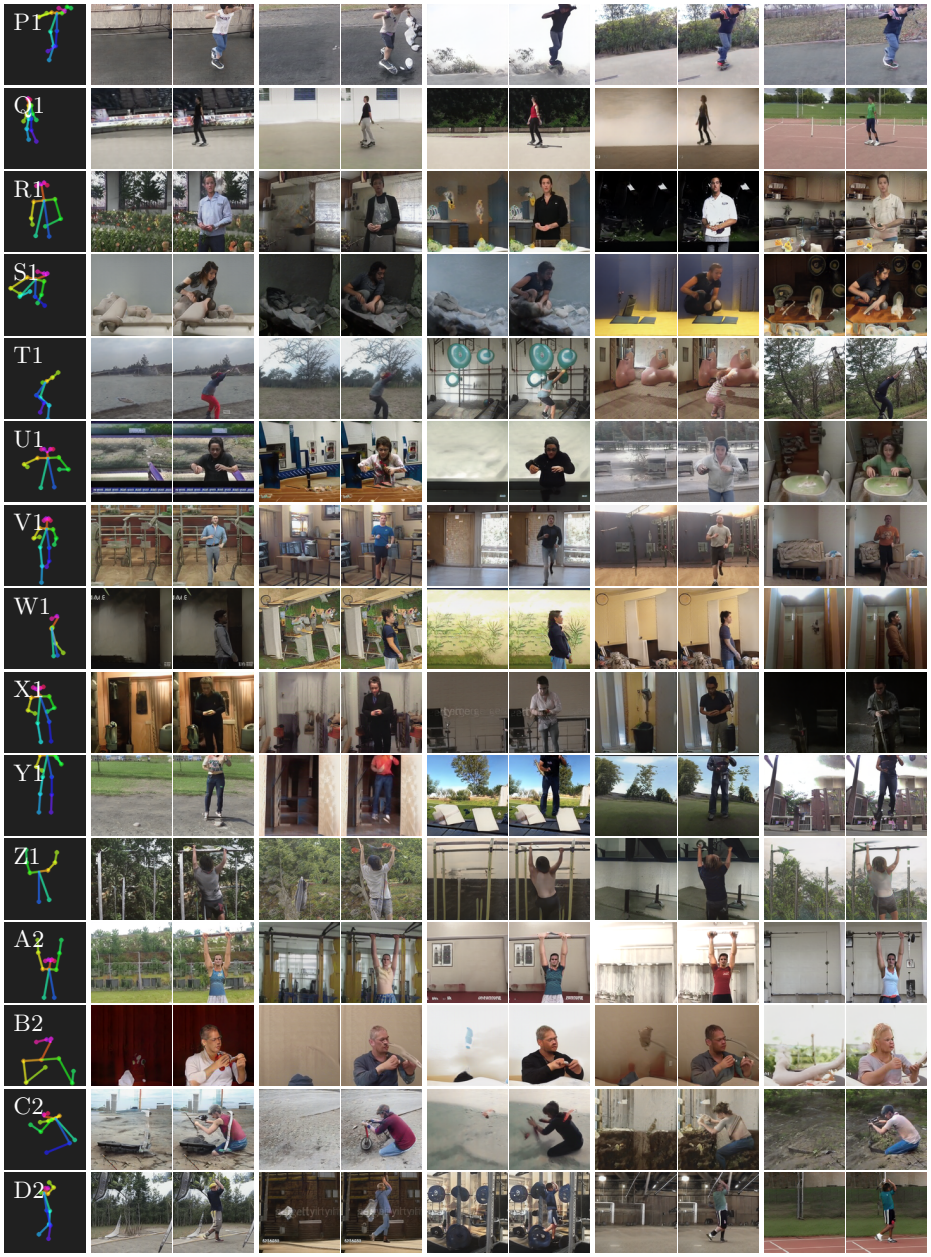


Fig. 4: **Random samples.**

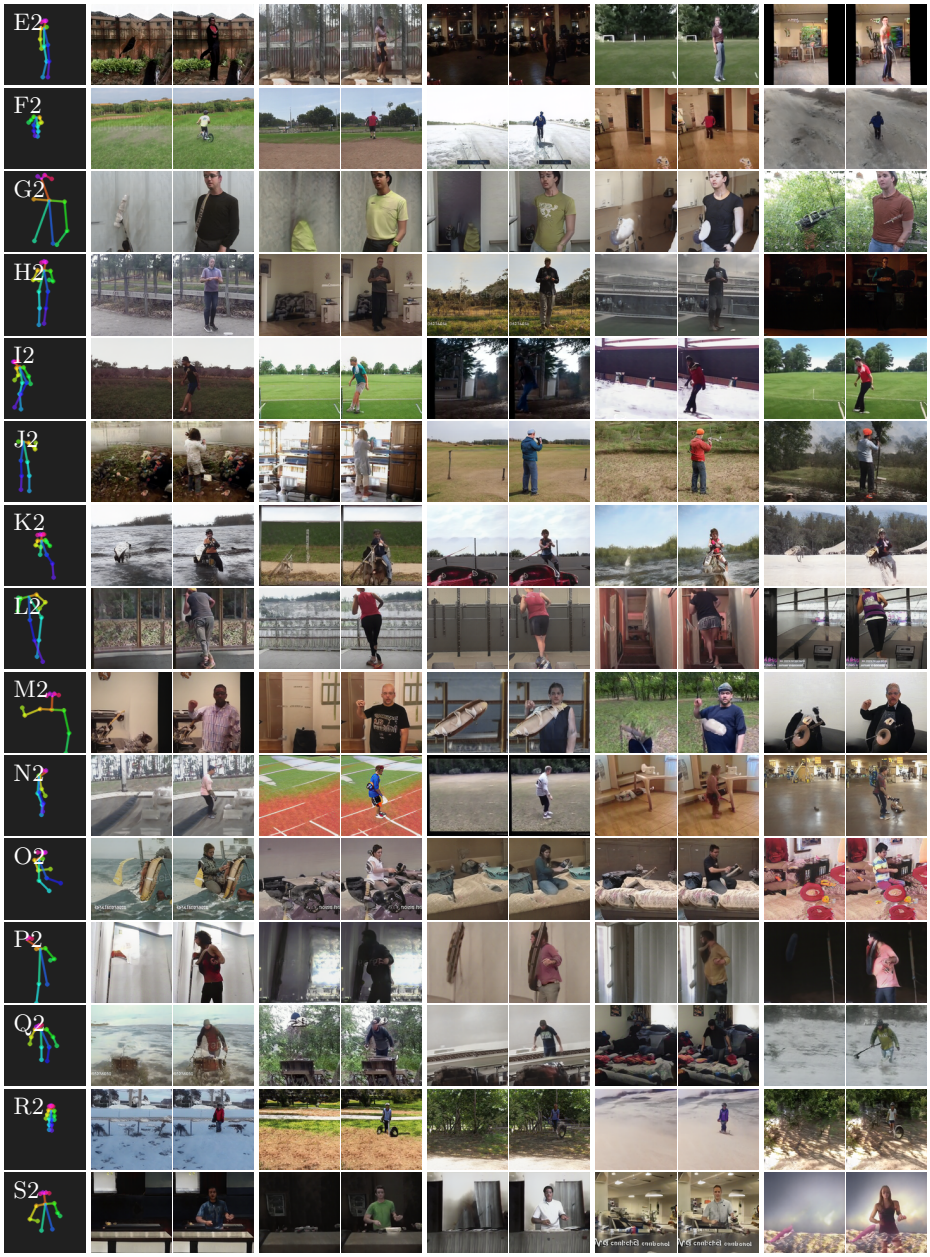


Fig. 5: **Random samples.**

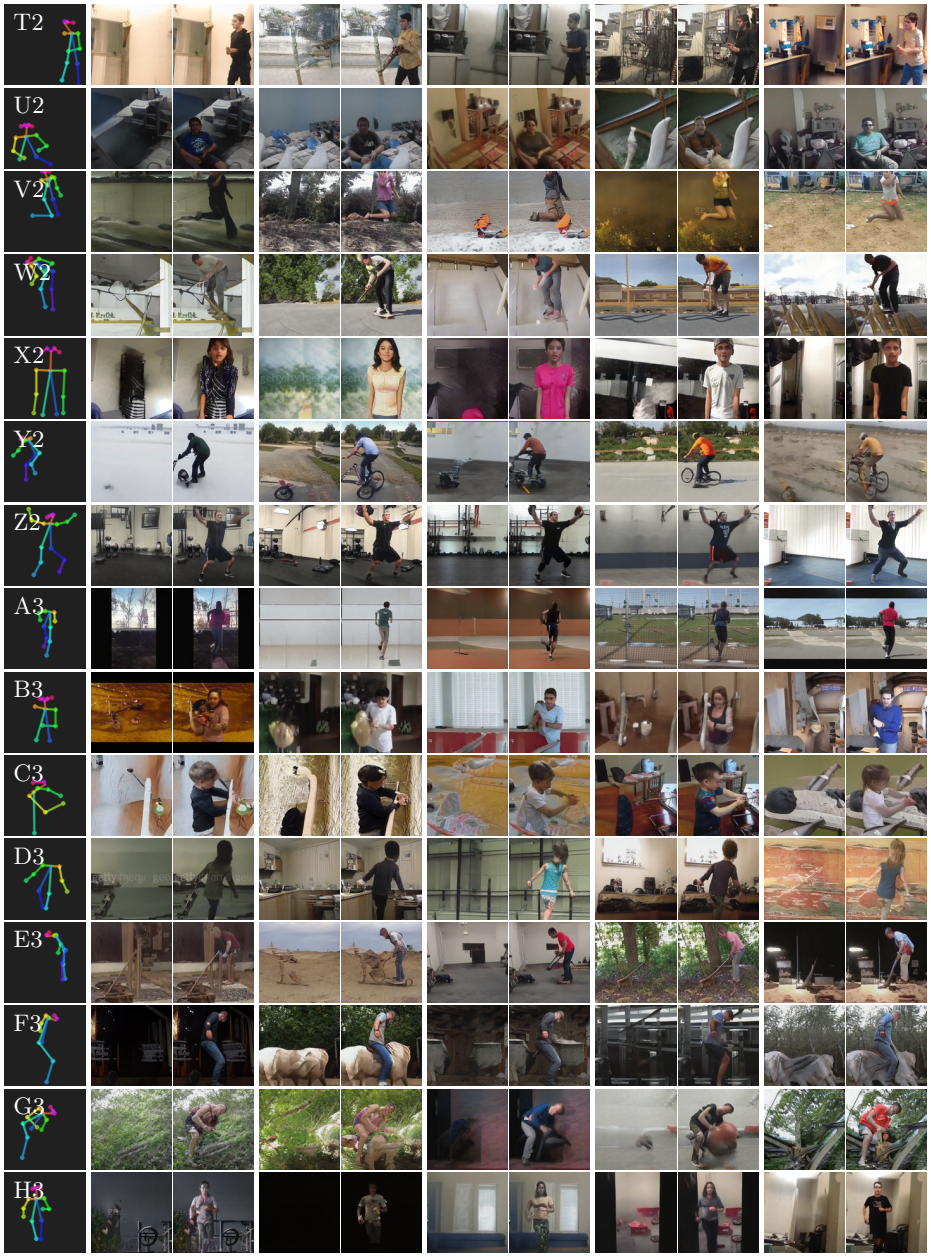


Fig. 6: **Random samples.**

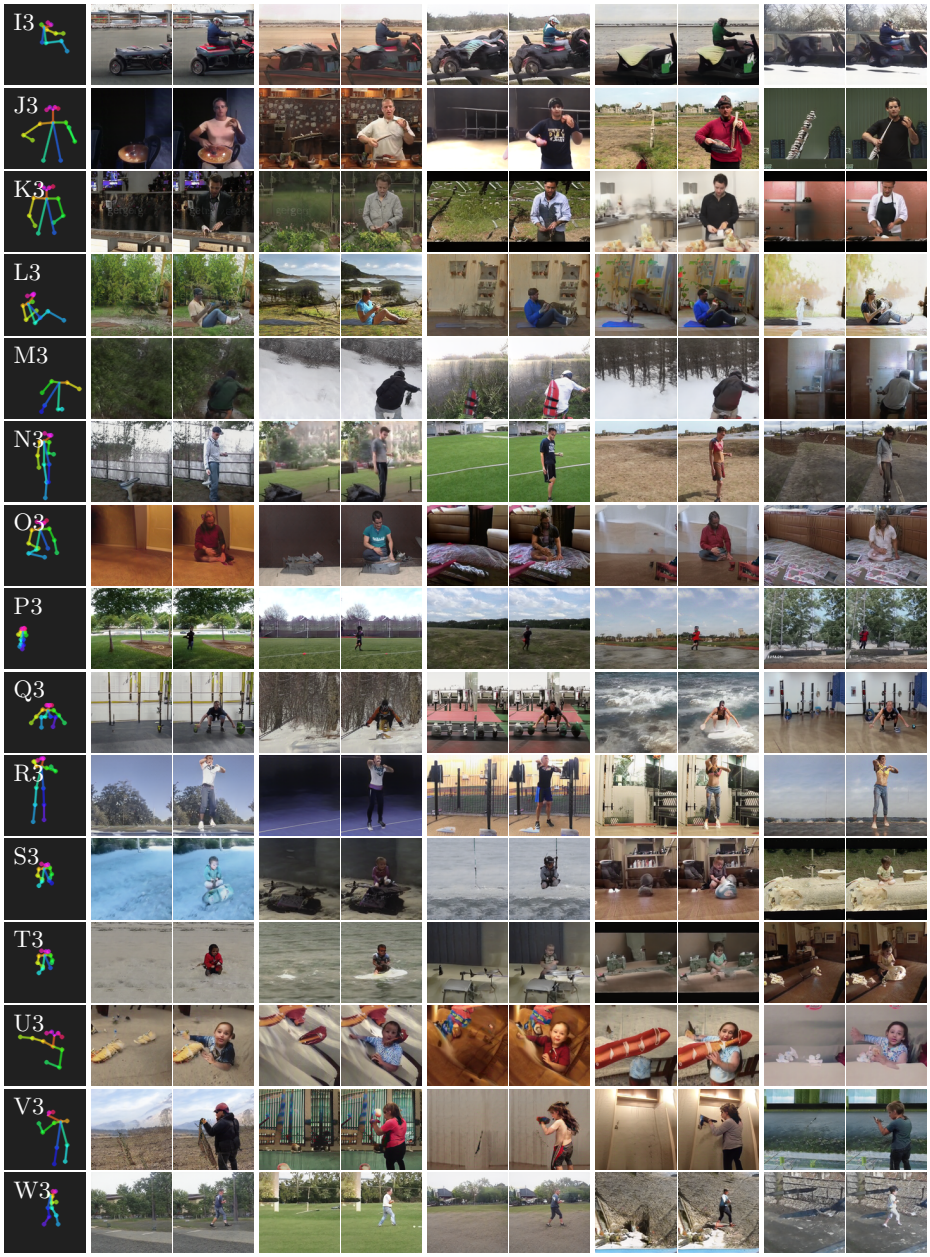


Fig. 7: **Random samples.**

References

1. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan: How to embed images into the stylegan latent space? In: ICCV. pp. 4431–4440 (2019), https://doi.org/10.1109/ICCV.2019.00453 5
2. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. pp. 3686–3693 (2014) 1, 2, 3
3. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: realtime multi-person 2d pose estimation using part affinity fields. IEEE transactions on pattern analysis and machine intelligence **43**(1), 172–186 (2019) 1, 2
4. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7291–7299 (2017) 1, 2
5. Carreira, J., Noland, E., Hillier, C., Zisserman, A.: A short note on the kinetics-700 human action dataset. arXiv preprint arXiv:1907.06987 (2019) 2
6. Diba, A., Fayyaz, M., Sharma, V., Paluri, M., Gall, J., Stiefelhagen, R., Van Gool, L.: Large scale holistic video understanding. In: European Conference on Computer Vision. pp. 593–610. Springer (2020) 1, 2
7. Duchon, C.E.: Lanczos filtering in one and two dimensions. Journal of Applied Meteorology and Climatology **18**(8), 1016 – 1022 (1979). https://doi.org/10.1175/1520-0450(1979)018¡1016:LFIOAT¿2.0.CO;2, https://journals.ametsoc.org/view/journals/apme/18/8/1520-0450_1979_018_1016_lfioat_2_0_co_2.xml 1
8. Epstein, D., Chen, B., Vondrick, C.: Oops! predicting unintentional action in video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 919–929 (2020) 1, 2, 3
9. Fouhey, D.F., Kuo, W.c., Efros, A.A., Malik, J.: From lifestyle vlogs to everyday interactions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4991–5000 (2018) 2, 3
10. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. arXiv preprint arXiv:1406.2661 (2014) 3
11. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017) 1
12. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016) 5
13. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5614–5623 (2019) 2
14. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=Hk99zCeAb 3
15. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. arXiv preprint arXiv:2006.06676 (2020) 3
16. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition. pp. 8110–8119 (2020) 3, 5
17. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017) 2
18. Kingma, D.P., Ba, J.: Adam: A methodfor stochastic optimization. In: International Conference onLearning Representations (ICLR) (2015) 5
19. Mescheder, L., Nowozin, S., Geiger, A.: Which training methods for gans do actually converge? In: International Conference on Machine Learning (ICML) (2018) 3
20. Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S.A., Yan, T., Brown, L., Fan, Q., Gutfreund, D., Vondrick, C., et al.: Moments in time dataset: one million videos for event understanding. IEEE transactions on pattern analysis and machine intelligence **42**(2), 502–508 (2019) 1, 2
21. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: International Conference on Machine Learning. pp. 1060–1069. PMLR (2016) 4
22. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: European Conference on Computer Vision. pp. 510–526. Springer (2016) 2
23. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2 (2019) 1
24. Xu, N., Yang, L., Fan, Y., Yang, J., Yue, D., Liang, Y., Price, B., Cohen, S., Huang, T.: Youtube-vos: Sequence-to-sequence video object segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 585–601 (2018) 2, 3
25. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018) 5
26. Zhang, W., Zhu, M., Derpanis, K.G.: From actemes to action: A strongly-supervised representation for detailed action understanding. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2248–2255 (2013) 2, 3
27. Zhao, S., Liu, Z., Lin, J., Zhu, J.Y., Han, S.: Differentiable augmentation for data-efficient gan training. arXiv preprint arXiv:2006.10738 (2020) 3