Supplementary Material for Aligning Videos in Space and Time

Senthil Purushwalkam*[1], Tian Ye*[1], Saurabh Gupta[2], and Abhinav Gupta[1,3]

[1] Carnegie Mellon University
[2] University of Illinois at Urbana-Champaign
[3] Facebook AI Research

1 Additional experimental details

Penn Action Dataset provides the training and testing splits. We further split the training data into 90 percent training data and 10 percent validation data for each action class. We use the validation set to tune the hyperparameters. We used a weight decay of 0.00001 with the Adam optimizer. We also observed that batch-normalizing the activations of the added linear layers improves performance both qualitatively and quantitatively.

2 Additional Visualization for Penn Action results

We attach additional visualizations for video retrieval and spatio-temporal alignment in Figure 1, Figure 2, Figure 3, Figure 4 and Figure 5.

Fig. 1: Visualization of results baseball pitch and baseball swing: Given a test video with tracks (shown in row 1, tracks denoted by colored markers), we retrieve and align a video from the training set from the same class (shown in row2). Retrieved videos are aligned in space (correspondence across videos is shown by the same colored marker), and in time (frames below one another are aligned). We see that our model is able to learn temporal and spatial correspondence. These visualizations are sampled *randomly*.

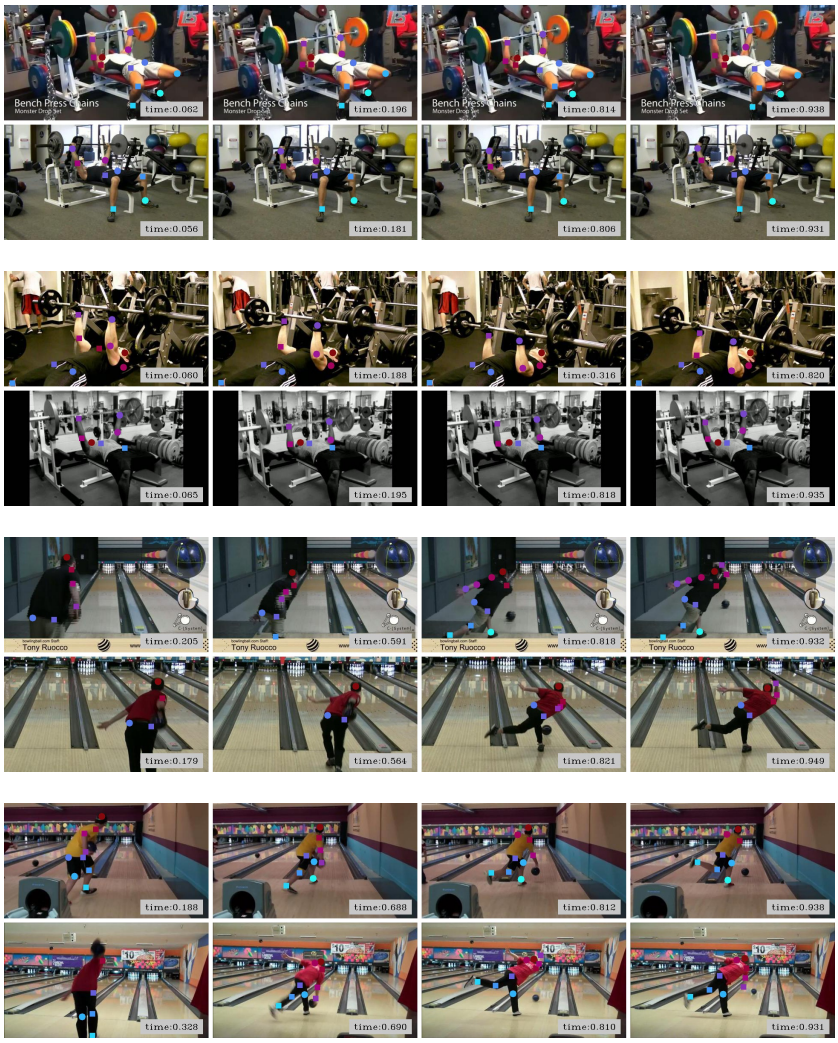

Fig. 2: Visualization of results bench press and bowling: Given a test video with tracks (shown in row 1, tracks denoted by colored markers), we retrieve and align a video from the training set from the same class (shown in row2). Retrieved videos are aligned in space (correspondence across videos is shown by the same colored marker), and in time (frames below one another are aligned). We see that our model is able to learn temporal and spatial correspondence. These visualizations are sampled *randomly*.

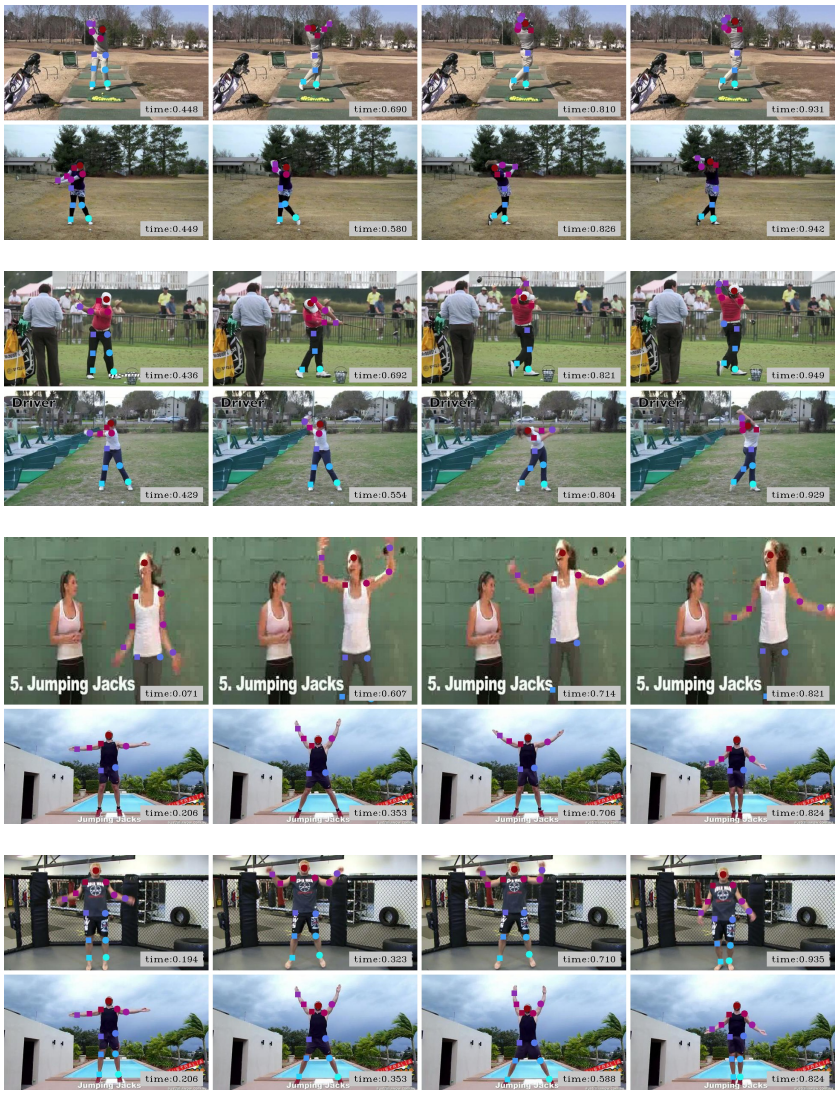

Fig. 3: Visualization of results golf swing and jumping jacks: Given a test video with tracks (shown in row 1, tracks denoted by colored markers), we retrieve and align a video from the training set from the same class (shown in row2). Retrieved videos are aligned in space (correspondence across videos is shown by the same colored marker), and in time (frames below one another are aligned). We see that our model is able to learn temporal and spatial correspondence. These visualizations are sampled *randomly*.

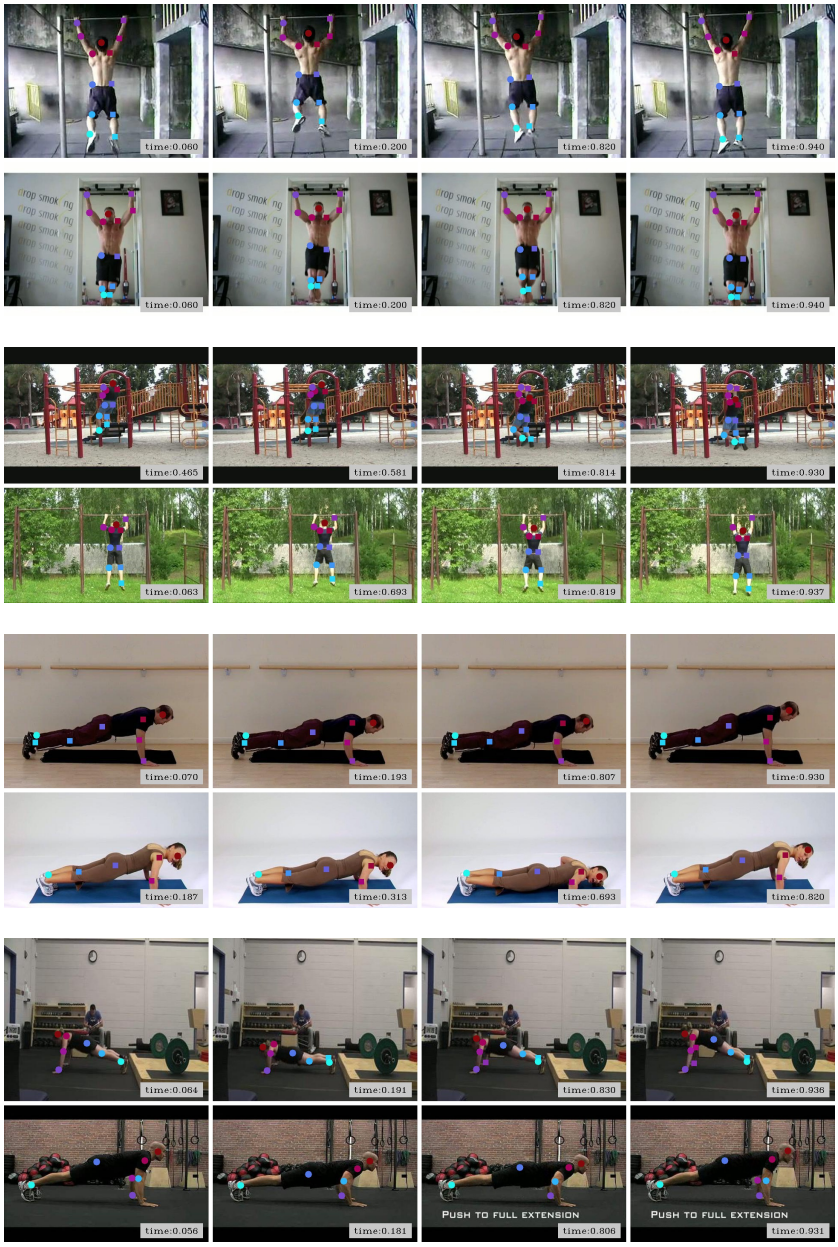

Fig. 4: Visualization of results for pullup and pushup: Given a test video with tracks (shown in row 1, tracks denoted by colored markers), we retrieve and align a video from the training set from the same class (shown in row2). Retrieved videos are aligned in space (correspondence across videos is shown by the same colored marker), and in time (frames below one another are aligned). We see that our model is able to learn temporal and spatial correspondence. These visualizations are sampled *randomly*.

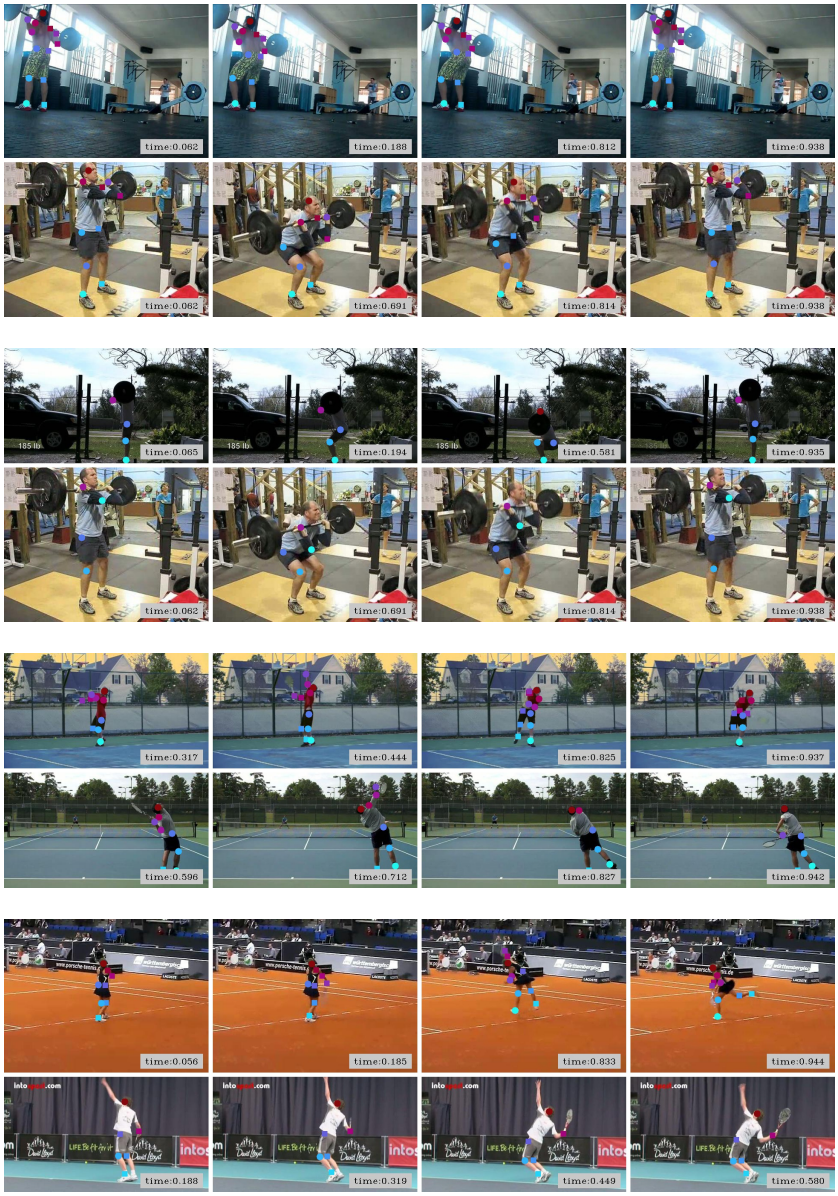

Fig. 5: Visualization of results for squat and tennis serve: Given a test video with tracks (shown in row 1, tracks denoted by colored markers), we retrieve and align a video from the training set from the same class (shown in row2). Retrieved videos are aligned in space (correspondence across videos is shown by the same colored marker), and in time (frames below one another are aligned). We see that our model is able to learn temporal and spatial correspondence. These visualizations are sampled *randomly*.

3 Visualization for EPIC Kitchen Dataset results

We present qualitative visualizations for video retrieval and alignment on the EPIC Kitchen dataset in Figure 6.

Fig. 6: Qualitative Results on Epic Kitchens [5]: We show qualitative examples of retrieval and temporal alignment from the Epic Kitchen Dataset, based on the similarity metric learned by our model.

References

1. Agrawal, P., Carreira, J., Malik, J.: Learning to see by moving. In: ICCV (2015)
2. Alayrac, J.B., Sivic, J., Laptev, I., Lacoste-Julien, S.: Joint discovery of object states and manipulation actions. In: ICCV (2017)
3. BK, P.H., Schunck, B.G.: Determining optical flow. Artificial intelligence **17**(1–3) (1981)
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017)
5. Damen, D., Doughty, H., Maria Farinella, G., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Scaling egocentric vision: The epic-kitchens dataset. In: ECCV (2018) 7
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR (2009)
7. Denton, E., Fergus, R.: Stochastic video generation with a learned prior. In: ICML (2018)
8. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: ICCV (2015)
9. Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., Zisserman, A.: Temporal cycle-consistency learning. In: CVPR (2019)
10. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: ICCV (2019)
11. Fouhey, D.F., Kuo, W., Efros, A.A., Malik, J.: From lifestyle vlogs to everyday interactions. In: CVPR (2018)
12. Garro, V., Fusiello, A., Savarese, S.: Label transfer exploiting three-dimensional structure for semantic segmentation. In: Proceedings of the 6th International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications (2013)
13. Girdhar, R., Carreira, J., Doersch, C., Zisserman, A.: Video action transformer network. In: CVPR (2019)
14. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: CVPR (2017)
15. Goyal, R., Kahou, S.E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The "something something" video database for learning and evaluating visual common sense. In: ICCV. vol. 1 (2017)
16. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., et al.: Ava: A video dataset of spatio-temporally localized atomic visual actions. In: CVPR (2018)
17. Ham, B., Cho, M., Schmid, C., Ponce, J.: Proposal flow. In: CVPR (2016)
18. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV (2017)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
20. Jayaraman, D., Grauman, K.: Learning image representations tied to ego-motion. In: ICCV (2015)
21. Kalal, Z., Mikolajczyk, K., Matas, J.: Forward-backward error: Automatic detection of tracking failures. In: ICPR (2010)
22. 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)

23. Kim, J., Liu, C., Sha, F., Grauman, K.: Deformable spatial pyramid matching for fast dense correspondences. In: CVPR (2013)
24. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
25. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
26. Lee, A.X., Zhang, R., Ebert, F., Abbeel, P., Finn, C., Levine, S.: Stochastic adversarial video prediction (2018)
27. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV (2014)
28. Liu, C., Yuen, J., Torralba, A.: Nonparametric scene parsing: Label transfer via dense scene alignment. In: CVPR (2009)
29. Liu, C., Yuen, J., Torralba, A.: Sift flow: Dense correspondence across scenes and its applications. TPAMI **33**(5) (2010)
30. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60**(2) (2004)
31. Lucas, B.D., Kanade, T., et al.: An iterative image registration technique with an application to stereo vision (1981)
32. Mémin, E., Pérez, P.: Dense estimation and object-based segmentation of the optical flow with robust techniques. IEEE Transactions on Image Processing **7**(5) (1998)
33. Misra, I., Zitnick, C.L., Hebert, M.: Shuffle and learn: unsupervised learning using temporal order verification. In: ECCV. Springer (2016)
34. Nagarajan, T., Feichtenhofer, C., Grauman, K.: Grounded human-object interaction hotspots from video. In: ICCV (2019)
35. Pathak, D., Girshick, R., Dollár, P., Darrell, T., Hariharan, B.: Learning features by watching objects move. In: CVPR (2017)
36. Rocco, I., Arandjelović, R., Sivic, J.: End-to-end weakly-supervised semantic alignment. In: CVPR (2018)
37. Rubinstein, M., Joulin, A., Kopf, J., Liu, C.: Unsupervised joint object discovery and segmentation in internet images. In: CVPR (2013)
38. Sermanet, P., Lynch, C., Chebotar, Y., Hsu, J., Jang, E., Schaal, S., Levine, S., Brain, G.: Time-contrastive networks: Self-supervised learning from video. In: ICRA. Pouring dataset licensed under (CC BY 4.0). (2018)
39. Sethi, I.K., Jain, R.: Finding trajectories of feature points in a monocular image sequence. TPAMI (1) (1987)
40. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: ECCV (2016)
41. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NIPS (2014)
42. Singh, S., Gupta, A., Efros, A.A.: Unsupervised discovery of mid-level discriminative patches. In: ECCV. Springer (2012)
43. Sun, D., Roth, S., Black, M.J.: Secrets of optical flow estimation and their principles. In: CVPR (2010)
44. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: CVPR (2018)
45. Tran, D., Bourdev, L.D., Fergus, R., Torresani, L., Paluri, M.: C3D: generic features for video analysis. CoRR, abs/1412.0767 **2**(7) (2014)
46. Varol, G., Laptev, I., Schmid, C.: Long-term temporal convolutions for action recognition. TPAMI **40**(6) (2017)

47. Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: BMVC (2009)
48. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: ECCV. Springer (2016)
49. Wang, X., Farhadi, A., Gupta, A.: Actions ~ transformations. In: CVPR (2016)
50. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR (2018)
51. Wang, X., Gupta, A.: Unsupervised learning of visual representations using videos. In: ICCV (2015)
52. Wang, X., Jabri, A., Efros, A.A.: Learning correspondence from the cycle-consistency of time. In: CVPR (2019)
53. Wu, C.Y., Feichtenhofer, C., Fan, H., He, K., Krahenbuhl, P., Girshick, R.: Long-term feature banks for detailed video understanding. In: CVPR (2019)
54. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: ECCV (2018)
55. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. TPAMI (2012)
56. Yeung, S., Russakovsky, O., Jin, N., Andriluka, M., Mori, G., Fei-Fei, L.: Every moment counts: Dense detailed labeling of actions in complex videos. IJCV **126**(2-4) (2018)
57. Zhang, H., Xiao, J., Quan, L.: Supervised label transfer for semantic segmentation of street scenes. In: ECCV. Springer (2010)
58. Zhang, R., Isola, P., Efros, A.A.: Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In: CVPR (2017)
59. Zhang, W., Zhu, M., Derpanis, K.G.: From actemes to action: A strongly-supervised representation for detailed action understanding. In: ICCV (2013)
60. Zhou, T., Jae Lee, Y., Yu, S.X., Efros, A.A.: Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences. In: CVPR (2015)
61. Zhou, T., Krahenbuhl, P., Aubry, M., Huang, Q., Efros, A.A.: Learning dense correspondence via 3d-guided cycle consistency. In: CVPR (2016)
62. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)