Appendix

A. More Implementation Details

Code will be made available.

A.1. Synthetic datasets

For fair comparison, we directly used the experimental setting and the evaluation protocol in [12]. The 2D ring has 8 Gaussian grid (std=0.05) equally distributed on a circle with $r = 2$, while the 2D grid has 5x5 Gaussian grid (std=0.1) equally distributed in a 8x8 square.

A.2. RGB Hand datasets

Our experiments were conducted on three datasets. First, we augmented a large-scale synthetic hand dataset. Then, we tested our method on two real datasets.

GANerated Hands [7]. This dataset is a synthetic dataset with 332k RGB images for hands. Those images were translated from SynthHands [8] via cycle consistency [2]. We collected frequently interacted objects from COCO [5] and inserted it onto the images without objects to form a new set of with-object data. In this way, we could get the object masks with visibility annotation. Note that currently there is no dataset with visibility annotation available ([13] does not release their used dataset). For the experiments, we equally (50/50) split the train/test set. The final dataset consists of 143k images for training.

Stereo Hand Benchmark [14]. This dataset consists of 12 sequences including 36k rgb images without hand-object interactions. For the experiments, we used the conventional split in [16] for direct comparison, where 10 sequences with 30k images were used for training and the rest were used for testing.

First-person Hand Action Benchmark (FPHAB) [1]. This large-scale dataset has 1200 sequences. We used the 280 sequences on hand-object interaction with 6-DOF object annotations. Specifically, we used 227 sequences with 17k images for training and 53 sequences with 4k images for testing.

A.3. Network architecture

For the experiments on synthetic Gaussian datasets, image generation and text-to-image translation, we directly borrow the architecture of the baseline methods [12, 9, 10] respectively for fair comparison. For the task of hand pose estimation, our network architecture is illustrated in Figure 1. We used the same design protocol as [7, 11], where 2D heatmap was first estimated to guide the 3D joint predictions. When computing the l_2 distance, only visible joints were considered.

A.4. Differentiable 2D projection

We used the projected 2D heatmap in the image-pose GAN formulation. However, it is worth noting that simply projecting and transferring the predicted 3D joints into 2D heatmap is non-differentiable. To get the meaningful gradient, we employed the differentiable image sampling technique in [3]. In this way, we could reparametrize the 2D heatmap with respect to the predicted 3D pose.

A.5. More detailed experimental settings

Image-to-image translation. Our implementation is mainly based on the official code[1] provided by [15] where we adapted the same architectures for the generator and the discriminator. Compared to the original BicycleGAN [15] implementation, we removed the image encoder which maps the RGB images to a coding space for reparametrizing the Gaussian distribution and replaced the instance normalization with spectral normalization [6]. We followed the same train/val/test split setting as [15]. We trained our model with a batch size of 8 for 325 epochs. We used $\alpha = 0.8$ for normalized diversity loss and 2.0 as L1 reconstruction weighted factor. During training, we randomly sampled 6 codes from the latent space to computed the diversity loss. Our initial learning rate was 2e-4 which was decayed by 0.1 for every 200 epochs. For quantitative evaluation, we randomly selected 100 images from validation set and sampled 38 outputs for each input image.

B. More Qualitative Results

We show more qualitative results on multimodal hand pose estimation from RGB images in Figure 2.

[1] https://github.com/junyanz/BicycleGAN

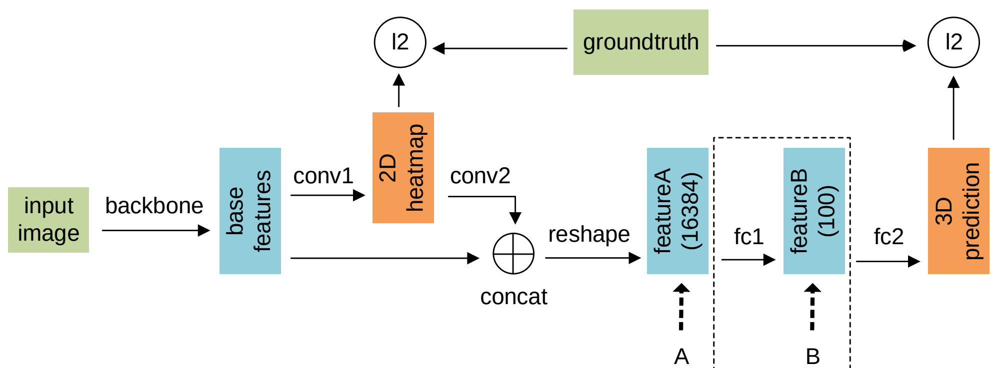

Figure 1: Network architecture for hand pose estimation. For the **backbone**, we directly borrow the architecture from [7]. Following the design protocol of [7, 11], we use the extracted **base features** to first reconstruct the **2D heatmap**. Then the predicted heatmap is concatenated with the base features. 'A' and 'B' denote where the noise vector $z \in \mathbb{R}^{10}$ is included via concatenation in 'Ours+' and 'Ours' respectively. The part in the dashed line, which is a bottleneck structure, is not used in 'Ours+'. Only visible joints contribute to the l_2 distance for both the 2D heatmap and 3D predictions. Specifically, the input image is sized 128x128. 'conv1' denotes one stride-1 conv layer and two stride-2 deconv layers. 'conv2' denotes two stride-2 conv layers. Both 'fc1' and 'fc2' denotes two sequential fc layers. The final 3D predictions has a dimension of 21x3=63.

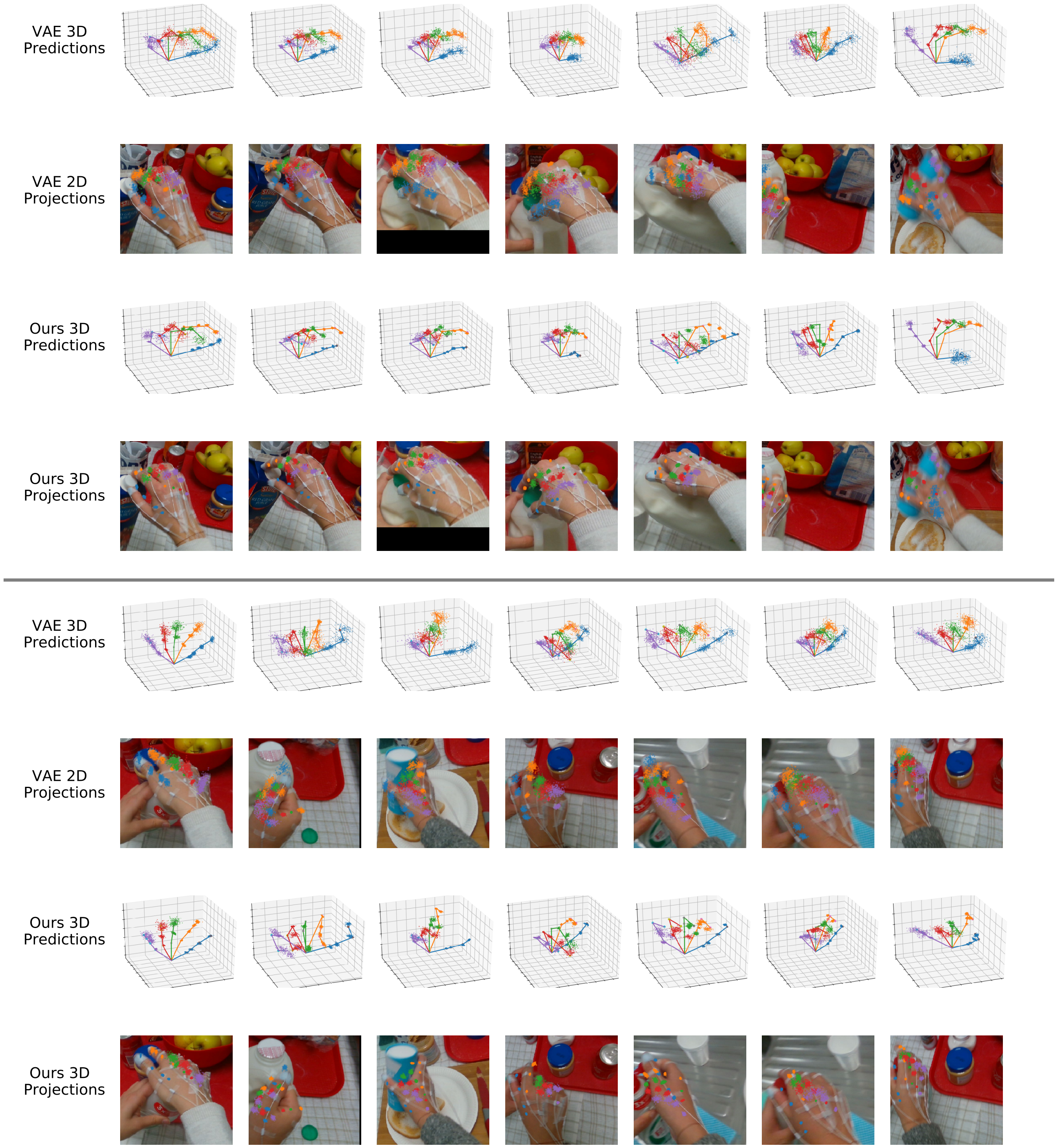

Figure 2: More qualitative comparison between VAE [4] and our method on 3D hand predictions and its projections on 2D image (better viewed when zoomed in).

References

[1] G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *CVPR*, 2018. 1

[2] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 1

[3] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *NIPS*, pages 2017–2025, 2015. 1

[4] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[5] 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 *ECCV*, pages 740–755. Springer, 2014. 1

[6] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018. 1

[7] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, and C. Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *CVPR*, pages 49–59, 2018. 1, 2

[8] F. Mueller, D. Mehta, O. Sotnychenko, S. Sridhar, D. Casas, and C. Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *ICCV*, volume 10, 2017. 1

[9] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 1

[10] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In *ICML*, pages 1060–1069, 2016. 1

[11] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, pages 4724–4732, 2016. 1, 2

[12] C. Xiao, P. Zhong, and C. Zheng. Bourgan: Generative networks with metric embeddings. In *NeurIPS*, 2018. 1

[13] Q. Ye and T.-K. Kim. Occlusion-aware hand pose estimation using hierarchical mixture density network. In *ECCV*, 2018. 1

[14] J. Zhang, J. Jiao, M. Chen, L. Qu, X. Xu, and Q. Yang. 3d hand pose tracking and estimation using stereo matching. *arXiv preprint arXiv:1610.07214*, 2016. 1

[15] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman. Toward multimodal image-to-image translation. In *NIPS*, pages 465–476, 2017. 1

[16] C. Zimmermann and T. Brox. Learning to estimate 3d hand pose from single rgb images. In *ICCV*, pages 4913–4921. IEEE, 2017. 1