Supplementary Material for
Structure-Aware Motion Transfer with Deformable Anchor Model

Jiale Tao[1*] Biao Wang[2] Borun Xu[1*] Tiezheng Ge[2] Yuning Jiang[2] Wen Li[1†] Lixin Duan[1]

[1]School of Computer Science and Engineering & Shenzhen Institute for Advanced Study, UESTC [2]Alibaba Group

{jialetao.std, liwenbnu, lxduan}@gmail.com, xbr_2017@std.uestc.edu.cn

{eric.wb, tiezheng.gtz, mengzhu.jyn}@alibaba-inc.com

In this supplementary material, we provide the implementation details of our proposed approach, more results on structure visualization, and further analysis on ablation results. We also provide the video results for comparison with state-of-the-art approaches, ablation study, and structure visualization in supplementary.mp4.

1. Implementation Details

We implement our model based on the released codes of [6, 7]. In particular, A U-Net [5] with skip connections is adopted as the image generator, and an Hourglass [4] network is utilized to predict motion and root anchors together with their affine transformation parameters. The flow mask estimator is also implemented by a U-Net. For the TaichiHD dataset, we further employ a background motion predictor to estimate the background motion similar to [7], which is implemented as the encoder part of a U-Net.

The inputs of the motion estimator, flow mask estimator and background motion predictor are all at the resolution of $4\times$ down-sampled input image similar to [6, 7]. The numbers of motion anchors and latent root anchor are set to 10 and 1 for all datasets. The intermediate anchors is set to 3 for the TaichiHD and Voxceleb1 datasets, and 4 for the MGIF dataset. The geometric transformation $\mathbf{T}$ (in Eqn. (11) of the main paper) is implemented as a random TPS (thin-plate spline [1]) transformation. And the input image is augmented under a random affine transformation.

We adopt Adam [3] optimizer with initial learning rate 0.0002 and decay it at the end of 60 and 90 training epochs by a factor $\alpha = 0.1$. The norm of the gradients of the motion estimator is clipped to 1 for stability following [2]. The batch size is set to 32 for 256×256 input resolution and 16 for 512×512. We train all datasets for 100 epochs using 4 Tesla V100 cards. Other implementation settings are exactly the same as those in [6].

Attention of Anchors: As formulated in Eqn. (14) of the main paper, an attention layer is designed to estimate the constraint relation between the intermediate anchors and motion anchors. Specifically, we take out the feature map $F^d \in \mathbb{R}^{C\times H\times W}$ of the driving frame of the last layer of the motion estimator, and bi-linearly sample the point features at the position of intermediate anchors ($z_i^d, i = 1, ..., I$) and motion anchors ($z_k^d, k = 1, ..., K$) respectively, denoted as $F_I^d \in \mathbb{R}^{C\times 1\times 1\times I}, F_K^d \in \mathbb{R}^{C\times 1\times 1\times K}$, which we further reshape to $\mathbb{R}^{I\times C}$ and $\mathbb{R}^{K\times C}$. We introduce $W_{Key}, W_{Que} \in \mathbb{R}^{C\times c}$, where $c = 64$, as the parameters of two fully connected layers, the attention weight ω is then computed as follow:

$$\omega = softmax\left(\frac{(F_I^d W_{Key})(F_K^d W_{Que})^T}{\sqrt{c}}\right) \tag{1}$$

where $\omega \in \mathbb{R}^{I\times K}$ and $\sum_k \omega_{ik} = 1, \forall i$. We further let $\omega_i \in \mathbb{R}^{1\times K}, \omega_k \in \mathbb{R}^{I\times 1}$ denote the i-th row and k-th column of ω.

Directly learning the attention weight ω via the HDAM loss $\mathcal{L}_{hdam}$ in Eqn. (14) of the main paper might cause a trivial solution that each row of the attention weight ω is activated by a single element only. To avoid this trivial solution and facilitate the better learning of ω, we propose the orthogonal loss and the completeness loss to constrain the rows and columns of ω:

$$\mathcal{L}_{ortho} = \sum_i \sum_{j\neq i} - \left\|\omega_i - \omega_j\right\|_1, \tag{2}$$

$$\mathcal{L}_{compl} = \sum_{k, \|\omega_k\|_1 < \delta} (\delta - \|\omega_k\|_1) \tag{3}$$

where δ is a predefined threshold and empirically set to 0.3 in our experiments. Intuitively, the orthogonal loss is designed to let different intermediate anchors control different motion anchors, and the completeness loss is to ensure that each motion anchor should be controlled by a larger weight than a give threshold attention value δ. These two losses are combined together with $\mathcal{L}_{hdam}$ and trained equally with those in Eqn. (15) of the main paper.

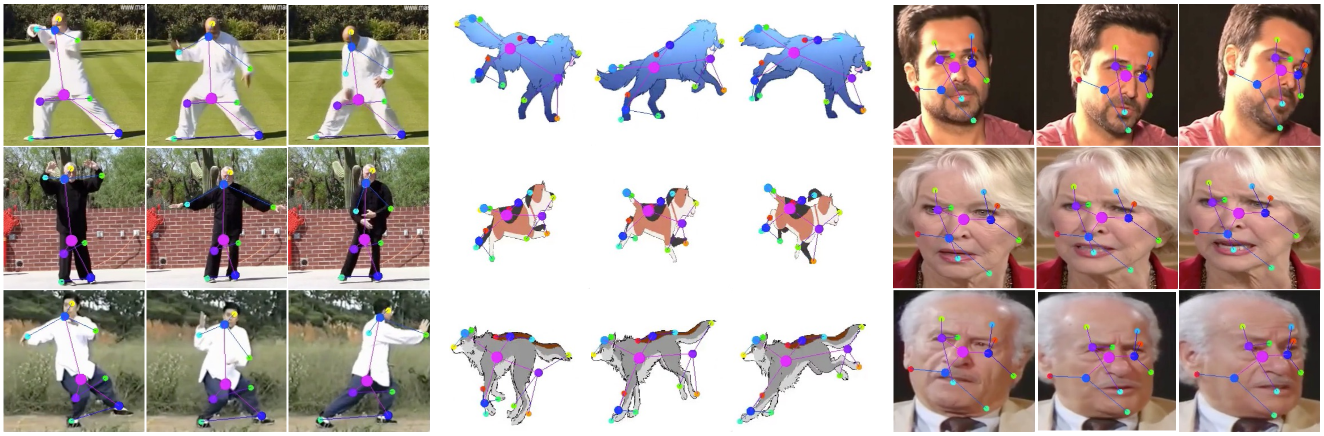

Figure S1. Additional structure visualizations on the three datasets.

Equivariance detection: Due to the lack of structural supervision in this task, a few of the learned motion anchors can be meaningless in some situations, which means it does not contributes to the optical flow estimation process (Eqn. (3) of the main paper) and the corresponding weight mask M tends to be zero at all spatial locations. To our findings, these motion anchors have a large equivariance loss defined in Eqn. (11) of the main paper, which means it tends to be predicted to a fixed coordinate location given any images. By contrast, a normal motion anchor is usually predicted to a fixed physical position given any images, and the predicted coordinate varies when motion occurs in the corresponding physical part. Given an empirically set value $\epsilon = 0.3$, we detect the abnormal motion anchors by judging if the equivariance loss of a motion anchor on the training set is larger than ϵ, and then remove those motion anchors in the attention process and in our qualitative results.

2. Additional Results on Structure Visualization

We provide more examples of the discovered object structure in Fig. S1. The video results for structure visualization are also provided. Similar to the findings in the main paper, we observe that the learned root anchor is generally located at the object centroid regardless of its identity or background; moreover, intermediate root anchors are often located at different local regions of an object, which enables them to capture more detailed motions of these object parts. This clearly demonstrate the effectiveness of our proposed HDAM approach on discovering the object structure when performing motion transfer.

3. Further Analysis on Ablation Results

As can be seen from Figure 1 and Figure 6 of our main paper, the intermediate anchors learned on the TaichiHD dataset often overlap with a motion anchor. We analyze this in a simple situation: assuming there are two motion anchors k_1, k_2 controlled by a intermediate anchor i with equal attention weights, then according to Eqn. (1) and Eqn. (8) of the main paper, the constraint loss can be computed as follow:

$$\mathcal{L}_{k_1,k_2 \leftarrow i} = \left\| z^s_{k_1} - \mathcal{T}_i\left(z^d_{k_1}\right)\right\|_2 + \left\| z^s_{k_2} - \mathcal{T}_i\left(z^d_{k_2}\right)\right\|_2 \quad (4)$$

We further assume that $\mathcal{L}_{k_1,k_2 \leftarrow i}$ is minimized to zero for simplicity, then we have:

$$\left\| z^s_{k_1} - \mathcal{T}_i\left(z^d_{k_1}\right)\right\|_2 = -\left\| z^s_{k_2} - \mathcal{T}_i\left(z^d_{k_2}\right)\right\|_2 \quad (5)$$

The meaning of Eqn. 5 is that the two motion anchors k_1 and k_2 is constrained by the affine transformation $\mathcal{T}_i$ of the intermediate anchor. Assuming the learned intermediate anchor is overlapped with the motion anchor k_1, which means:

$$z^d_i = z^d_{k_1}, T_i(z^d_{k_1}) = T_i(z^d_i) = z^s_i = z^s_{k_1} \quad (6)$$

According to Eqn. 6, the left side of Eqn. 5 is zero. Moreover, by using Eqn. 6 and Eqn. (7) of the main paper, we can compute $\mathcal{T}_i\left(z^d_{k_2}\right)$ as follow:

$$\mathcal{T}_i\left(z^d_{k_2}\right) = z^s_{k_1} + \theta_i(z^d_{k_2} - z^d_{k_1}) \quad (7)$$

Summarizing Eqn. 5, Eqn. 6 and Eqn. 7 we can obtain:

$$\theta_i(z^d_{k_2} - z^d_{k_1}) = z^s_{k_2} - z^s_{k_1} \quad (8)$$

The Eqn. 8 indicates that the motion anchors k_1 and k_2 in the source and driving image are explicitly constrained by the affine transformation θ_i of the intermediate anchor. This explains that, the overlapping between intermediate anchors and motion anchors, is reasonable and can ease the learning process. We take a simple situation for analysis, while in those complicated and non-overlapping situations, the constraint relations between different motion anchors are implicitly implemented through intermediate anchors.

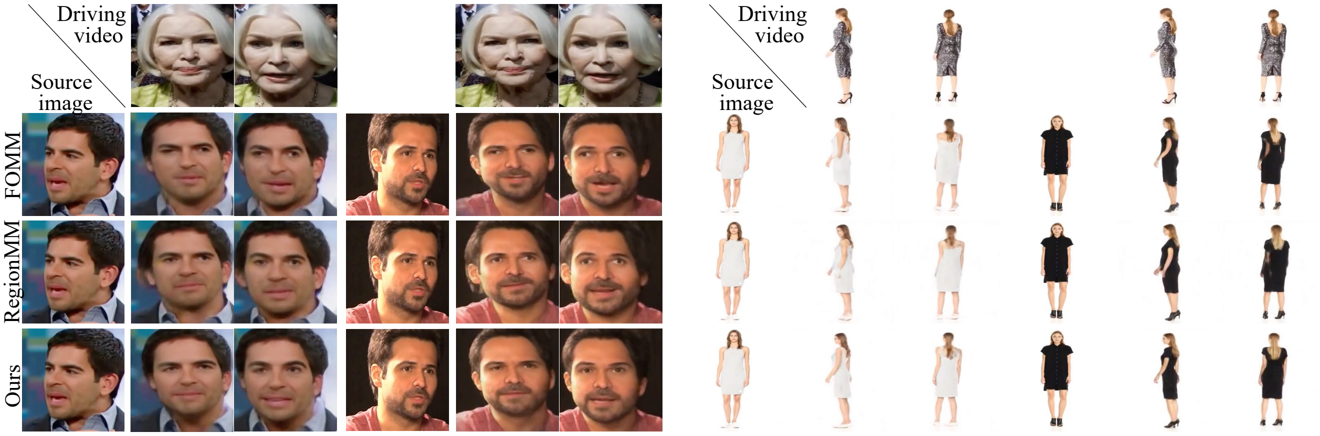

Figure S2. Qualitative comparison on the Voxceleb1 and FashionVideo datasets.

4. Video results

We give video results in the provided video file, including qualitative comparison, ablation study and structure visualizations. For analysis, we present the generated key frames on the Voxceleb1 and FashionVideo datasets in Fig. S2. As can be seen from the left side of the figure, generated videos using our method better preserves the structure of source faces, while FOMM [6] and RegionMM [7] suffers from more distortions in the synthesized face structures. And as shown in the right side of the figure, on the Fashion dataset, our methods generates more stable arm motions when the model turns around. These results suggest that our HDAM model generally constrains the object structures well, which enables it to synthesize more structure-stable videos.

References

[1] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence*, 11(6):567–585, 1989. 1

[2] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Self-supervised learning of interpretable keypoints from unlabelled videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8787–8797, 2020. 1

[3] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 1

[4] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 1

[5] 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

[6] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Advances in Neural Information Processing Systems*, 2019. 1, 3

[7] Aliaksandr Siarohin, Oliver Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *CVPR*, 2021. 1, 3