VideoMotionCLIP: Animating Human Images Using Free Text

Amnon Levi, Guy Tevet, Amit Bermano

Abstract

We propose a novel pipeline for synthesizing human videos from one or more source images coupled with a descriptive free text description. Our method leverages the embedding of human motion into the CLIP disentangled latent space presented in MotionCLIP, to generate human poses from free text. We then use the motion and a learned transformation flow to animate the human in the source image, over an in-painted background. The result is a video of the photographed human performing the described action over the original background.

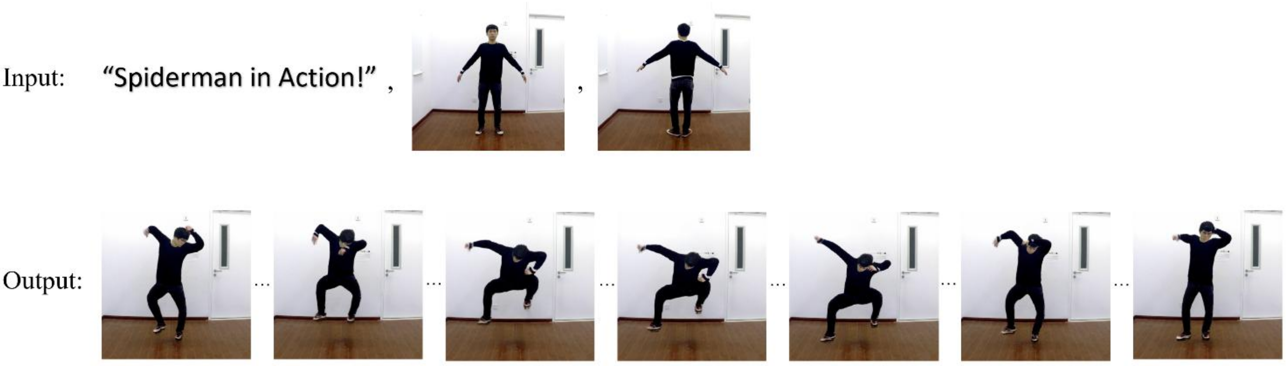

Fig. 1. Conditional text and some reference images result in a conditioned video.

1. Introduction

While image synthesis and editing advanced in both realism and expressiveness, video synthesis is still at a much earlier stage. Current popular methods require a reference video dictating the motion, paired with a source style [Zhu et al. 2017], a video [Chan et al. 2019], or a set of images [Liu et al. 2020].

In this work, we aim to remove the dependency on a reference video, by synthesizing the motion from free text leveraging the expressiveness and semantic properties of the CLIP latent space, using the generation method presented in MotionCLIP [Tevet et al. 2022].

Using a free text description to animate an existing image allows the synthesis of videos without the need of a talented animator creating the motions, or an actor, or a dancer performing the motions. Another benefit is quick, semantic, intentional modifications to the resulting video, by changing part(s) of the description given. for example, changing the word “jogging” to “running away”, will result in a different synthesized video, allowing for very short experimentation cycles.

2. Related Work

2.1 Style Editing

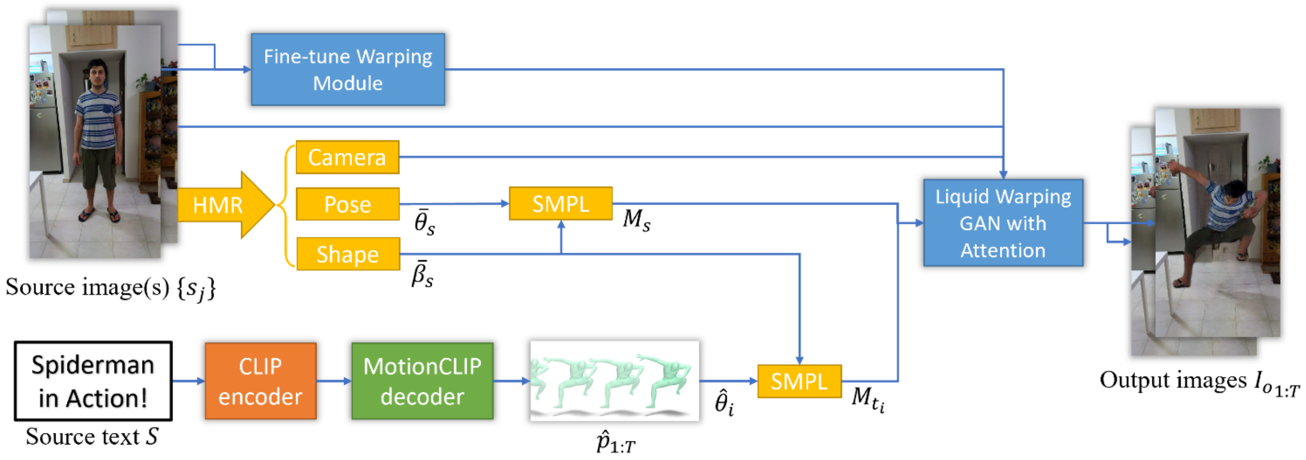

Fig. 2. The inference pipeline

One way to synthesize a video is by changing the style of an existing video. CycleGAN [Zhu et al. 2017] trained a Generative Adversarial Network (GAN) to transform, for example, horses into zebras, and then synthesized zebra videos from pre-existing horse videos.

2.2 Motion Transfer

A very popular method for video generation is the combination of a source, which will be animated, with some reference, which will dictate the motion.

First Order Motion [Siarohin et al. 2019] uses an image source along with a reference video. Their proposed method is learning to transform a frame to keypoints and their local affine transformations, as well as learning to interpolate the sparse motion to a dense one. This is a general method that is not limited to human motion but requires learning a new model for a given set of motion types.

Everybody Dance Now [Chan et al. 2019] uses a video source and a video reference and uses a GAN to learn to transform a pair of consecutive motions, to a pair of consecutive frames in the domain of the source video. This is then combined with pose estimation to transfer the motion from the reference video to the actor and environment of the source video. When further combined a face GAN, this work creates very impressive results, but it requires a video source, and long training times per source.

Neural Actor [Liu et al. 2021] also uses a video(s) source and a video reference. Their method first learns a 3D human mesh of the actor in the source video, and then uses pose estimation to extract the motion from the reference video. Both are then combined to render the source actor performing the reference motion.

2.3 CLIP aided Methods

CLIP [Radford et al. 2021] is a model coupling images and text strings in a joint deep latent space which was trained on hundreds of millions of image-description pairs. Among other things, this expressive, semantic coupling opened the door to a flood of image generation using a text description, including, but not limited to, DALL-E 2 [Ramesh et al. 2022] and Imagen [Saharia et al. 2022].

AvatarCLIP [Hong et al. 2022] presents a text to animated avatar generation. The suggested method implements 3 learned processes – coarse body shape generation, shape sculpting and texture generation, and motion generation. The first two steps guided by CLIP, while the third step takes inspiration from

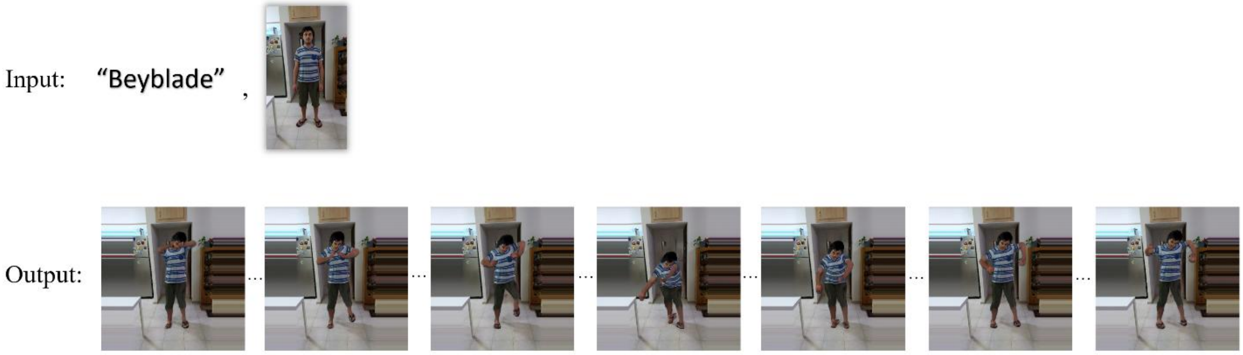

Fig. 3. An in-the-wild image with an out-of-domain phrase. Note how the expressiveness of CLIP helps capturing an irregular motion associated with a pop-culture reference.

ACTOR [Petrovich et al. 2021] and uses a motion VAE.

3. Method

3.1 Preliminaries

MotionCLIP [Tevet et al. 2022] uses a transformer VAE to embed motion into a latent space aligned with CLIP's latent space. This is obtained by cosine distance losses between their encoding of the motion and CLIP's encoding of the motion description text and a random rendered frame from the motion, in addition to the standard VAE reconstruction loss. The result is a text and motion coupling with an expressive latent space, obtained with the relatively limited existing data, by leveraging CLIP's existing knowledge. This allows for out of domain motion generation from free text, as well as semantic editing using latent-space arithmetic.

Liquid Warping GAN with Attention [Liu et al. 2020], uses an image(s) source and a video reference. Their method creates a transformation flow from one or more images $T_{si\to t}$ and fine tunes it using a GAN combining 3 internal CNNs. G_{SID} takes a correspondence map C_{si} and a foreground image I_{si}^{ft} , and reconstructs the foreground and mask. G_{TSF} does a similar thing, but with a synthesize foreground image I_t^{syn} and a target correspondence map C_t. Both use Attention Liquid Warping Blocks that take $T_{si\to t}$ as an input. The final CNN G_{BG} takes a masked background image I_{bg} , and synthesizes the full background image I_{bg}.

This pretrained GAN architecture is the fine-tuned for each source using the one or more source images, to achieve a more accurate result. The resulting transformation flow is then combined with pose estimation of the reference video to render the source actor performing the reference motions in their original environment.

3.2 Pipeline Overview

Using the method presented in MotionClip [Tevet et al. 2022], we encode the input text string S to a latent representation z_p using the transformer encoder presented in CLIP [Radford et al. 2021]. We then decode it using the transformer decoder presented in MotionCLIP, with z_p as the key and value, and the positional encoding of $1{:}T$ as the query, to obtain the motion data $p_{1:T}$.

We represent motion sequences using the SMPL body model [Loper et al. 2015]. A sequence of length T denoted $p_{1:T}$ such that $p_i \in \mathbb{R}^{24\times6}$ defines orientations in 6D representation [Zhou et al. 2019] for global body orientation and 23 SMPL joints, at the i^{th} frame. We denote the position and pose data of each frame i as $\theta_i \in \mathbb{R}^{24\times3}$.

For each source image s_i, we estimate the SMPL shape parameters of the source model $\overline{\beta}_{s_i} \in \mathbb{R}^{10}$ which was estimated using the Body Mesh Recovery Module presented in Liquid Warping Gan with Attention using HMR [Kanazawa et al. 2018] [Kolotouros et al. 2019]. We then use the method presented in Liquid Warping GAN with Attention [Liu et al. 2020] for the source images input $\{I_{si}\}$ to obtain the transformation flow T.

We combine the source shape parameters $\overline{\beta}_s$ and the generated position and pose parameters for each frame θ_i to obtain the SMPL mesh $M_{ti} = M \;\; \theta_i, \overline{\beta}_s$. For each frame i, we infer the motion imitation pipeline of Liquid Warping GAN with Attention using T and M_{ti} to render the frame I_{oi}. The frames $I_{o1:T}$ are then compiled into an output video V_o. See figure 2 for an overview of the complete pipeline.

4. Results

We implemented the presented pipeline and have successfully run it using images and text phrases shown in the preliminary papers (see figure 1). We also tested using in-the-wild images with novel, out-of-domain phrases which are semantically alien to the phrases present in the training data, with similar success (see figure 3).

5. Conclusions

We presented a video generation/ images animation method from one or more images and a single phrase, which leverages the expressiveness of CLIP to create motions from any text description and use them to animate an actor in a scene.

Our method has several limitations. Existing methods for using few or a single image to animate a person in a scene are still in their infancy, with the resulting videos, while relatively temporally stable, containing many artifacts. Due to the current size of labeled data, text-to-motion techniques are not yet expressive enough, and many phrases result in generic movement. Additionally, while the generated motions may sometimes appear coherent with a stick model or an SMPL rendered model, they fall into the uncanny valley when used on a real human in a realistic scenario.

There are many more research opportunities in improving the results by way of training with more data, improving individual modules used in the method, or altering it altogether.

References

Zhu, J.Y., Park, T., Isola, P. and Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision (pp. 2223-2232).

Chan, C., Ginosar, S., Zhou, T. and Efros, A.A., 2019. Everybody dance now. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 5933-5942).

Liu, W., Piao, Z., Tu, Z., Luo, W., Ma, L. and Gao, S., 2021. Liquid warping gan with attention: A unified framework for human image synthesis. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Tevet, G., Gordon, B., Hertz, A., Bermano, A.H. and Cohen-Or, D., 2022. MotionCLIP: Exposing Human Motion Generation to CLIP Space. arXiv preprint arXiv:2203.08063.

Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E. and Sebe, N., 2019. First order motion model for image animation. Advances in Neural Information Processing Systems, 32.

Liu, L., Habermann, M., Rudnev, V., Sarkar, K., Gu, J. and Theobalt, C., 2021. Neural actor: Neural free-view synthesis of human actors with pose control. ACM Transactions on Graphics (TOG), 40(6), pp.1-16.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. and Krueger, G., 2021, July. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (pp. 8748-8763). PMLR.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. and Chen, M., 2022. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G. and Salimans, T., 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. arXiv preprint arXiv:2205.11487.

Hong, F., Zhang, M., Pan, L., Cai, Z., Yang, L. and Liu, Z., 2022. AvatarCLIP: Zero-Shot Text-Driven Generation and Animation of 3D Avatars. arXiv preprint arXiv:2205.08535.

Petrovich, M., Black, M.J. and Varol, G., 2021. Action-conditioned 3d human motion synthesis with transformer vae. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10985-10995).

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G. and Black, M.J., 2015. SMPL: A skinned multi-person linear model. ACM transactions on graphics (TOG), 34(6), pp.1-16.

Zhou, Y., Barnes, C., Lu, J., Yang, J. and Li, H., 2019. On the continuity of rotation representations in neural networks. In Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5745-5753).

Kanazawa, A., Black, M.J., Jacobs, D.W. and Malik, J., 2018. End-to-end recovery of human shape and pose. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7122-7131).

Kolotouros, N., Pavlakos, G., Black, M.J. and Daniilidis, K., 2019. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2252-2261).