Project Proposal: Multi-Object Motion Transfer

Written by Nikita Lockshin (NL2668), Xipeng Xie (XX2297), LianFeng Li (LL3226)

Previous works in "Animating Arbitrary Objects via Deep Motion Transfer" and its reference paper numbers marked [#] have been summarized as the following.

- Generative Adversarial Networks (GAN) and Variation Autoencoders (VAE) need labels for training.
- Deep generative models can transfer motion patterns or a facial expression but need ground-truth to pre-train models to learn a specific object, so Deep generative models are not good for arbitrary objects.
- Spatio-temporal networks: VGAN [42] and TGAN [30] can generate frames from video but in poor quality.
- Recurrent Neural Networks (RNN) with adversarial training framework: (CMM-Net) [45] and MoGoGAN [39] can get high quality frames.
- [34, 26, 23] directly predict raw pixel values in future frames; [13, 40, 2] propose learning the transformation which maps pixels in given frames to future frames; [41] uses LSTM to predict the motion of landmarks, and then generates images from them.
- [7, 38] are only applied for specific domains, and [5] [49] are for facial, so they are not good for challenging situations.
- Recycle GAN is only used for 2 specific domains, but the authors want to depict unknown object in driving video.
- [9][43] propose spatio-temporal to get motion transfer.
- [46] uses X2Face to modify input image to follow the motion pattern in another modality, such as audio.

From unsupervised landmark discovery in [20, 47], the **Monkey-Net algorithm** detects motion by getting keypoints' locations' differences between 2 continuous frames in driving video (no labels and self-supervised), and then modifies the source image according to the landmarks extracted from the driving frame. The framework can be split into:

- Achieving sparse keypoints in the source image and video frame.
- Dense Motion prediction network translates them into motion heatmaps.
- Motion Transfer network recombines the source image and dense motion heatmap to be a target frame.

The paper uses following **datasets** to achieve **authors' goal** of getting an animated video from the source image and using motion patterns from driving video for an arbitrary object.

- UvA-Nemo [1]: pre-processing in [45], aligning face in OpenFace library [1] before re-sizing to 64x64, each video lasting 32 frames, 1110 videos for training, and 124 videos for testing.
- Tai-Chi [39]: pre-processing in [39], resizing to 64x64, lasting 32 to 100 frames, 3288 videos for training, and 822 videos for testing.
- BAIR robot in [11]

We will utilize the datasets of UvA-Nemo and Tai-Chi, and **our project goal** is to:

- Tier 1: Get fully understanding of the original code and how it works, and try to use it in other domain, such as making an animal show a human's impression.
- Tier 2: Reconstruct the network to make it possible to animate several objects in the source image.

We will use the same **evaluation criteria** as the paper. Authors use the following metrics to carry out an ablation study and to show that they beat previous works, most notably X2Face, unsupervised MoCoGAN, unsupervised SV2P, and CMM-Net.

- L_1: attempt to reconstruct the input video and evaluate the result against it.
- Average Keypoint Distance (AKD): use external keypoint detectors for ground truth and the generated video and measure the average distance between them.
- Missing Keypoint Rate (MKR): the percentage of keypoints present in the ground truth but not detected in the generated video.
- Average Euclidian Rate (AED): a feature representation metric employed in [12] to test the ground truth against the generated video.
- Frechet Inception Distance (FID) [18]: evaluate quality of individual frames.