Replacing a Person in a Video with Another

Dhanesh Pamnani Jiyoon Park Shri Ishwaryaa S V

1. Project Idea: What is the Idea?

We propose a model that can replace a character (called target from here on) in the video with another character (called ego from here on). Given a video containing the target, we extract the subject from the background and fill up the background either using diffusion or by using the information from the other frames (explained below in the down-scoping subsection). We extract the position of the target in each frame as well as estimate the pose to obtain a stick figure. We use deformable GANs to then transform our ego person to that obtained pose of the target person for each frame. This new pose of the ego person can then be put in the background of the original target video for each frame. We can also apply some kind of post-processing to try and replicate the lighting on the ego person as per the video lighting.

1.1. Down-scoping

For the initial implementation of the idea, we will start with videos shot with a fixed camera. The background of the video would also be static (i.e. a room) instead of a dynamic background with various components (i.e. traffic). These initial assumptions of the background will allow us to use the background information from the other frames to fill the background of the current frame. Using videos with a fixed camera view will also help in understanding and replicating the scale of the target and the ego person and make the replacement task easier. We start with the target making limited motions of its limbs in the front view facing the camera so that we have a simpler starting point. Eventually, we would want the target to move across the scene and interact with objects or go through occlusions. We also aim to process a dynamic background or an animated background with animated characters in the future.

2. Motivation: Why is it important and relevant? Why should we care?

This project aims to replace a human in a video with another human. This transports the human onto another environment or scene. This is useful in replacing stunt doubles in movie scenes with the real actor/actress. It can help social media influencers advertise products creatively. They could place themselves in any catchy/historical video while promoting products such as clothes, jewelry, wearable technology, etc. In addition, this can enable people to live vicariously through the videos that they are placed in. The project can be viewed as an extension of photoshop to videos.

3. Prior Work: Briefly mention related works that are relevant to your idea.

The paper 'Progressive Pose Attention for Person Image Generation', proposes a new generative adversarial network for pose transfer, i.e., transferring the pose of a given person to a target pose. It uses the DeepFashion data set for the training.
The paper 'Appearance and Pose-Conditioned Human Image Generation using Deformable GANs' addresses the problem of generating person images conditioned on both pose and appearance information. The paper '2d human pose estimation: New benchmark and state of the art analysis.' is also a great resource.
Single Person Pose Estimation has been done using several methods. The tree-structured graphical models use the spatial relationship between the joints and the non-tree-structured models are used to allow exceptions such as occlusions. Convolutional Neural Networks (CNNs) have also been used for body pose estimation.

4. Project Idea: Why does your idea make sense intuitively?

People's desire to experience more without engaging in the whole experience has triggered the acceleration of Virtual Reality and Augmented Reality development. We are now in a generation where people are satisfied enough with an adventure by only meeting the visual sensing requirements. We propose that replacing a character in a video with yourself will give people a sense of contentment and the feeling of experiencing the scene. There is existing software that replaces your face in videos, but none that incorporates the full motion of a person. Our project aims to continue and expand the work to replace the entire human body.

5. Project Idea: How does it relate to prior work in the area?

Extensive work has been done to replace a target person's face with an ego person to make it seem like the ego person is talking and making such facial expressions. There have been many works done to learn a person's speaking patterns and use them to generate speech data that that person did not say. We extend this flow of research to replace an entire human body with another.

6. Experiments: What evaluations are you aiming for?

The generated video should have the ego person replicate the pose of the target person. The background should be filled in and not have black spots. We understand that there would be limitations due to computation and also limitations in using deformable GANs that we are not aware of at this point, but want the produced frames to look as realistic as possible. The generated video should seem realistic enough for the ego subject to look at and feel like they were actually in that scene. We aim to give an experience to the ego subject without making that person go do the task in that environment.

7. Baselines: What all baselines are you thinking of?

The baselines for this project are diffusion models and deformable GANs as we will be extracting the person from the background and using human pose estimation to obtain the target poses for the ego person.

References

1. Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In Proc. CVPR, 2014
2. Zhen Zhu1, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, Xiang Bai. Progressive Pose Attention Transfer for Person Image Generation. In CVPR, 2019
3. Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, Alexei A. Efros. Context Encoders: Feature Learning by Inpainting. CoRR, 2016
4. Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, Bryan Catanzaro. Image Inpainting for Irregular Holes Using Partial Convolutions. ECCV 2018
5. Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, Thomas S. Huang. Generative Image Inpainting with Contextual Attention. CVPR 2018