Deep Learned Face Swapping in Feature Film Production

Jonathan Swartz
Wētā FX
Wellington, New Zealand
jswartz@wetafx.co.nz

Sean Walker
Wētā FX
Wellington, New Zealand
seanwalker@wetafx.co.nz

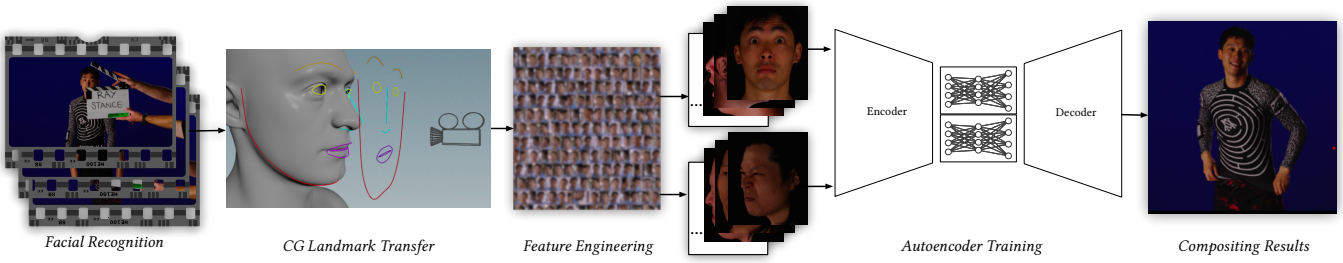

Figure 1: A schematic depicting our approach. From left to right: faces are identified with facial recognition; landmarks are transfered from CG assets; features in the data are emphasized for qualities relevant for a shot; datasets are fed to the autoencoder during training; and finally results are composited. Note: The persons featured are not the actors in the film.

ABSTRACT

In visual effects for film, replacement of stunt performers' facial likeness for their doubled actor counterparts using traditional computer graphics methods is a multi-stage, labor intensive task. Recently, deep learning techniques have made a compelling argument to train neural networks to learn how to take an image of a person's face and convincingly infer a rendered image of a second person's face with a previously unseen perspective, pose and lighting environment.

A novel method is discussed for bringing deep neural network face swapping to feature film production which utilizes facial recognition for the discovery of training data. Our method further innovates in the area of utilizing traditional CG assets for informing some of the shortcomings of ML techniques. Connected with a technique for feature engineering during training dataset assembly, our *Face Fabrication System* enables Wētā FX to deliver final picture quality for use in production.

ACM Reference Format:
Jonathan Swartz and Sean Walker. 2022. Deep Learned Face Swapping in Feature Film Production. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Talks (SIGGRAPH '22 Talks), August 07-11, 2022.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3532836.3536271

1 INTRODUCTION

The film *Shang-Chi and the Legend of the Ten Rings* contains many shots which required martial arts performances by trained stunt performers. In order to retain the authenticity of the performances desired by the production, Wētā FX was employed to produce face replacements of the stunt performers for their doubled actors in these shots. The quantity of face replacements and artistic requirements lead us to examine recent work in face swapping using deep learning techniques.

Since early work on approaches to face swapping using deep convolutional neural networks [Korshunova et al. 2017], a number of software projects have proliferated offering casual users practical implementations for extracting faces from images, assembling training sets, training networks and rendering finished results [Deepfakes 2021], [Perov et al. 2020]. After evaluating several of these approaches for our specific needs and performing experiments with data acquired during a test shoot using feature-film quality camera equipment, we decided our aim was to bring a derivative approach based on these projects into production.

2 METHODOLOGY

Our approach is deployed in a classic visual effects workflow and, as such, we wanted to utilize as much existing data as possible to augment the deep learning techniques while also delivering the rendered faces into a traditional compositing pipeline for as much flexibility in finishing as possible. Additionally, there is no time allotted in vfx production for atypical tasks, such as exploring and labelling training data. Therefore, it is important to identify opportunities to develop secondary systems which assist with deep learning tasks without adding manual labour overhead. To those ends, we have developed novel methods for utilizing our facial neural rendering techniques in feature film production including:

2.1 Facial Recognition for Data Harvesting

Given that the methods we are utilizing rely on training a model using pairs of images from a dataset of face images of two identities, building such datasets of our actors and stunt performers is of the

SIGGRAPH '22 Talks, August 07-11, 2022, Vancouver, BC, Canada

ACM ISBN 978-1-4503-9371-3/22/08.
https://doi.org/10.1145/3532836.3536271

utmost importance to the approach. *Shang-Chi* provided additional challenges because it was photographed during the SARS-CoV-2 pandemic which restricted the feasability of photographing additional reference specifically for our technique. Therefore, to build our training data, we decided to make use of heterogeneous sources of photographed faces such as witness cameras placed around the shooting set, footage captured for prinicipal photography across all sequences in the show, and action tests or reference which the studio had shot for other purposes.

This set of source types presented us with hundreds of individual source media, each of which contained potential images of faces for our subjects as well as many images of people who were not our subjects. Having a person to sort footage and hand-label the faces in the footage was not feasible within the constraints of our project. This lead us to use ML techniques in facial recognition to automate much of this process.

Using pretrained deep neural networks (DNNs) for both facial bounds detection [Chi et al. 2018] and landmark identification [Yang et al. 2020], we were able to take each piece of footage, recognize regions of frames with faces and produce images for all faces wherein key facial landmarks are consistently aligned. Having consistently aligned face images allowed us to infer the identities of these faces from a model we trained to identify the actors and stunt performers of interest. This facial identity classification model was built by training a support-vector machine to classify the identities of our actors based on the 512-dimensional feature vectors produced by an Inception-ResNet-v1 network pretrained on VGGFace2[Schroff et al. 2015]. The results of detecting, landmarking and identifying these faces were curated in a Neo4j graph database which was custom deployed for this project to allow natural flexibility for how we represented and searched information about media, faces and identities. In production we found this combination of facial identification and graph-style databasing to be a very practical approach to process, identify and index over 1.1 million faces which could be flexibly searched for use in downstream machine learning tasks.

2.2 Facial Landmarks from 3D Models

For training and inference, our neural rendering network requires all face images to be aligned to a normalized space based on key semantic landmarks on the face. Our project targets usage in action-heavy shots where the stunt actors' faces are occluded, at glancing angles or otherwise obscured. We have found that many deep learning approaches to facial landmark identification have poor accuracy in exactly these situations. In these cases, our approach was to rely on rotomation being generated by the Wētā Matchmove department, independently of our process, for its conventional use by the Animation, FX and Lighting departments.

By relying on Wētā's facial models' adherence to a consistent UV mapping relative to facial features, we could lay out the landmarks once in generic UV-space and then obtain consistent world-space positions of the key facial landmarks across all human characters in every shot we would be working in. For shots where our DNN techniques failed to infer meaningful landmarks, we were able to remain robust by augmenting those results with landmarks derived from the traditional pipeline by projecting the UV-space landmarks into the camera's image space for cataloguing in our facial database.

2.3 Feature Engineering

To prioritize producing high fidelity facial imagery over generalization, we trained separate models to render imagery for specific shots. This strategy was motivated by a pragmatic desire to focus the training time and capacity of a model to learn features that would be most useful to achieve the effect on a given shot. In order to do this, we had to find methodical techniques to shape the distribution of training images we were choosing from our database.

One method we employed to make sure we were learning from a high quality set of images was to query the database for face images that were at least of a minimum resolution given the face scale in a target shot. In order to try and ensure that these face images were not blurry due to motion blur, we further refined these selections by computing an image-space measurement of the magnitude of movement of the face across neighouring frames and removing faces where a great deal of global movement relative to the frame correlates with heavily motion-blurred images. Given our face database contains the positions for facial landmarks for all frames and their association to people's identities determined during facial recognition, we can inexpensively compute this measurement for a given face based on surrounding frames of the same person's face.

Lastly, we refine our training data by only selecting faces with similar orientations to the faces in the target shot. We do so by first estimating the 3D orientation of all faces in the target shot by using the 2D face landmarks' correspondence to a set of canonical 3D landmark positions in order to perform pose estimation of the faces. Faces for the training set are then selected based upon a maximum tolerance for the distance between rotations of each candidate face's pose and those in the target shot.

3 CONCLUSION AND FUTURE WORK

Our work marks the first usage of a practical implementation for face swapping with DNNs at Wētā which has been used in production to deliver final images on *Shang-Chi*. We look forward to continuing to improve *Face Fabrication System* in the areas of artistic control, fidelity and flexibility. One area we hope to explore is the possibility of introducing synthetic training images made by Wētā's pipeline which can be acommpanied by a richer set of correlated information about the face, camera and lighting.

REFERENCES

Cheng Chi, Shifeng Zhang, Junliang Xing, Zhen Lei, Stan Z. Li, and Xudong Zou. 2018. Selective Refinement Network for High Performance Face Detection. *arXiv:1809.02693 [cs]* (Sept. 2018). http://arxiv.org/abs/1809.02693 arXiv: 1809.02693.

Github Deepfakes. 2021. *FaceSwap.* https://github.com/deepfakes/faceswap Accessed 2021-02-21.

I. Korshunova, W. Shi, J. Dambre, and L. Theis. 2017. Fast Face-Swap Using Convolutional Neural Networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 3697–3705. https://doi.org/10.1109/ICCV.2017.397 ISSN: 2380-7504.

Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, et al. 2020. DeepFaceLab: A simple, flexible and extensible face swapping framework. *arXiv:2005.05535 [cs, eess]* (May 2020). http://arxiv.org/abs/2005.05535

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A Unified Embedding for Face Recognition and Clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition*, 815–823. https://doi.org/10.1109/CVPR.2015.7298682

Jing Yang, Adrian Bulat, and Georgios Tzimiropoulos. 2020. FAN-Face: a Simple Orthogonal Improvement to Deep Face Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 07, 12621–12628. https://doi.org/10.1609/aaai.v34i07.6953