Deep Video Portraits

- Project Link
- This is a SIGGRAPH 2018 paper.
- This paper has 10 authors. The first is Hyeongwoo Kim. So it goes by [Kim et al. 2018].
- **Abstract**
  - Goal = reanimation of a human subject in a video.
  - Workhorse = a neural network that transforms synthetic renderings of a parametric face model to a photo-realistic image of a human.
  - To reanimate:
    - Extract pose and blendshape parameters from source video.
    - Render the video of the face model.
    - Feed the video to the neural network to get a human video.
  - The paper is able to transfer
    - 3D head position,
    - head rotation,
    - facial expression,
    - eye gaze,
    - and eye blinking.

Introduction

- Video portrait = video framed to show a person's head and upper body.
- Problem the paper tries to solve.
  - Input
    - Video of a *source* actor.
    - Video of a *target* actor.
  - Output
    - A video of the target actor mimicking the action of the source actor.
- The paper claims that, from the target video, it can synthesize:
  - head pose,
  - facial expression,
  - eye gaze
  - identity (to some extent)
  - and consistent background.
- The paper views the probem as a rendering-to-video translation task.
  - Input
    - Synthetic renderings of a controllable 3D face model.
    - Rendered eye gaze images.
  - Ouptut

        - Photo-realistic rendering of human having the pose of the synthetic rendering.
- The paper uses a deep neural network to accomplish the above task.
    - It has encoder-decoder architecture.
    - It takes a sliding window of frames as input (i.e. it operates on space-time) and tries to predict the next frame.
    - It is trained adversarially, and so is a conditional GAN.
- The paper claims that, while it cannot control the full upper body pose because it only tracks the face, the network can generate plausible upper body pose and background.
- It seems that the network has to be retrained for each new target video.

Background

Here we have a list of cited papers.

- **Monocular face reconstruction**
    - [Blanz and Vetter 1999] is the first paper that fits a 3D morphable model to a photo. This is the paper that people always cite (SIGGRAPH 1999).
    - [Blanz et al. 2004] uses the 3D morphable model to exchange faces in photo, mainly trying to solve the problem of pose and illumination difference (Eurographics 2004).
    - [Kemelmacher-Shlizerman 2013] constructs a morphable face model from pictures found on the Internet (ICCV 2013).
    - [Kemelmacher-Shlizerman et al. 2010] creates an image database of a person from photos found on the internet and uses database lookup to do face puppetry (ECCV 2010).
    - [[Roth et al. 2016]] (http://cvlab.cse.msu.edu/pdfs/Roth_Tong_Liu_CVPR16.pdf) creates high quality face model from internet photo collections by fitting a morphable model to the photos and refining it. It is not so clear whether the final model is still morphable or not (CVPR 2016).
    - [Cao et al. 2014a]: So called the "DDE algorithm." This is used to track blendshapes from RGB camera input in real time without needing to calibrate for a new user beforehand.
    - [Fyffe et al. 2014] creates photo-realistic animation by driving one or more high-res facial scans with a video of the same person (SIGGRAPH 2014).
    - [Garrido et al. 2016] constructs a morphable face model from a monocular video (SIGGRAPH 2016).
    - [Ichim et al. 2015] creates a personalized morphable face model from photographs taken from a mobile phone. The user has to take specific photograph/video captures, including specific sweep video and making specific facial expressions (SIGGRAPH 2015).

    - [Suwajanakorn et al. 2014] creates a 3D model of the face region from each frame of the input video. Does not use any morphable model (ECCV 2014).
    - [Theis et al. 2016] transfers motion of a source video to a target video sequence (CVPR 2016).
        - Track face meshes in source and target videos.
        - Transfer expression by mesh deformation.
        - Use target video of fine the appropriate frame to copy the mouth interior for inpainting.
        - Does not change head pose.
    - [Wu et al. 2016] presents a new face model that can tracks local deformations caused by external forces (skin being blown by wind, for example). (SIGGRAPH 2016)
    - [Booth et al. 2018] presents a morphable face model constructed from about 10,000 faces (IJCV 2018).
    - [Richardson et al. 2016] use deep learning to construct 3D face models from single images. Only constructs the shape. No appearance. Use synthetic data for supervised learning. (3DV 2016)
    - [Tewari et al. 2017] constructs 3D face models with geometry and reflectance from single images. Model can be trained end-to-end in an unsupervised manner from a collection of unconstrained photos. (ICCV 2017)
    - [Tran et al. 2017] regresses a 3D morphable model and texture from photos with deep learning (CVPR 2017).
    - [Cao et al. 2015]: a real-time facial performance capture system that captures both global shapes and local details from a single RGB camera (SIGGRAPH 2015).
    - [Richardson et al. 2017] an upgrade to the 2016 paper by the same first author. Uses two networks for coarse and fine geometry reconstruction. Train on synthetic data in a supervised learning setting first, followed by a unsupervised learning phase on unconstrained facial images.
    - [Sela et al. 2017] constructs a face model with a depth map to capture fine-scale geometric details.
- **Video-based facial reenactment**
    - [Suwajanakorn et al. 2015] constructs a morphable model from a person's photographs found on the Internet.
    - [Vlasic et al. 2005] solves the expresion transfer problem using the approach where we extract movement → render face mesh → merge back to target image to animate face.
    - [Dale et al. 2011] replace the face in a target video with another face from the source video. Requires tracking facial performance in both videos with morphable models.
    - [Garrido et al. 2014] is another face transfer system, but it does not rely on 3D models like Dale et al. Frames from the source videos are retrieved,

  warped, and merged into the target video. All using 2D approaches.
  - [Li et al. 2014] uses frame in the target video to choose which frame to use for each source frame. Has quite a complicated algorithm in addition to the lookup. (IEEE Transactions on Multimedia)
  - [[Wood et al. 2018]] redirects the eye gaze in videos by fitting a 3D eyeball video to the input image and then uses it to render the eye portion, which is composited to the original image. (Eurographics 2018)
  - [Theis et al. 2018] presents a real-time face capture system for a user who's wearing a VR goggle and uses the capture to reanimate a photo-realistic face. (SIGGRAPH 2018)
  - [Fried et al. 2016] uses image warping to change head pose and camera position of a portrait image.
- **Visual dubbing**
  - Visual dubbing is the task of generating mout movement of a target actor to match an audio track.
  - [Bregler et al. 1997] tracks mouth points and morphs. Possibly the first paper on the topic?
  - [Chang and Ezzat 2005] allows reanimating the mouth of an actor with a small video corpus. Creates a multidimensional morphable model (MMM) from a large corpus first and then tries to adapt it to a small corpus. (SCA 2005)
  - [Ezzat et al. 2002] creates a multidimensional morphable model (MMM) from a large corpus of a person's video saying specific syllables and uses the model to create animation of speech later. (SIGGRAPH 2002)
  - [Liu and Ostermann 2011] build a database of mouth images and perform lookups to create videos. (ICME 2011)
  - [Suwajanakorn et al. 2017] trains a network to synthesize high quality mouth movements from speech and composes it to a new target video of the same person. Needs a collection of videos of the target person. (SIGGRAPH 2017)
  - [Garrido et al. 2015] transfers mouth movement of the dubbing actor to the target video. Uses morphable model to capture and transfer movement.
- **Image-to-image translation**
  - [Mirza and Osindero 2014] is the original conditional GAN paper.
  - [Hinton and Salakhutdinov 2006] introduces the "autoencoder."
  - [Radford et al. 2016] introduces the DCGAN architecture.
  - [Chen and Koltun 2017] describes a conditional GAN that can generates a very high resolution (2048×1024) images by cascades of refinements. Also introduces the "feature matching" loss. (ICCV 2017)
  - [Wang et al. 2018] is another conditional GAN that can generates high resolution images. Has better image quality than Chen and Koltun's paper. (CVPR 2018)
  - [Zhu et al. 2017] is the CycleGAN paper. (ICCV 2017)

    - [Yi et al. 2017] is the DualGAN paper. It has the same goal as CycleGAN. Same overall approach but different architecture. (ICCV 2017)
    - [Liu et al. 2017] is the UNIT paper. Again, another paper that does the same thing as CycleGAN does. Maps images from two domains to the same latent space before translating to the other domain. (NIPS 2017)
    - [Taigman et al. 2017] another unsupervised image-translation network. However, this time there is a network that models what must remains the same when translating between domains. (ICLR 2017)
    - [Ganin et al. 2016] uses image translation for gaze manipulation. (ECCV 2016)
    - [Olszewski et al. 2017] animates a face from a single image by creating a 3D model. Uses dynamic per-frame texture to capture fine details such as wrinkles. Deep networks are used to generate these textures. (ICCV 2017)
    - [Lassnet et al. 2017] gives a generative model for generating pictures of full-body people in clothing. Generates a semantic segmentation first and then converts the segmentation to the output image (ICCV 2017).
    - [Ma et al. 2017] gives a conditional GAN that accepts and image of a person and the desired full body pose and outputs a new image of the person with that pose. (NIPS 2017)

Overview

- Both the source and target videos are tracked to get per-frame parameters of a controllable face model.
- The parameters represent the actors'
    - identity,
    - head pose,
    - expression,
    - eye gaze,
    - and scene lighting.
- Then, we transfer the relevant parameters (head pose, expression, eye gaze) from the source actor to the target actor.
- New renderings of the face model are then generated. The renderings include:
    - Normal color rendering.
    - Correspondence map.
    - Eye gaze images.
- The renderings above serve as conditioning inputs to the render-to-video translation network.
- To gain temporal stability, the renderings are fed in a sliding window fashion.

Monocular Face Reconstruction

- The source video is denoted by $\mathcal{V}^s = \{\mathcal{I}_f^s : f = 1, \cdots, N_s\}$ where N_s is the number of frames.

- Similarly, the target video is denoted by $\mathcal{V}^t = \{\mathcal{I}_f^t : f = 1, \dots, N_t\}$ where N_t is the number of frames.
- Let $\mathcal{P}^{\square} = \{\mathcal{P}_f^{\square} : f = 1, \dots, N_{\square}\}$ denote the face model parameters for each of the frames in the source/target video.
- The parameters include:
    - The rigid head pose $R^{\square} \in SO(3)$.
    - The translation $\mathbf{t}^{\square} \in \mathbb{R}^3$.
    - The facial identity coefficients:
        - The geometry coefficients $\boldsymbol{\alpha}^{\square} \in \mathbb{R}^{N_\alpha}$ where $N_\alpha = 80$.
        - The reflectance coefficients $\boldsymbol{\beta}^{\square} \in \mathbb{R}^{N_\beta}$ where $N_\beta = 80$.
    - The expression coefficients $\boldsymbol{\delta}^{\square} \in \mathbb{R}^{N_\delta}$ where $N_\delta = 64$.
    - The gaze directon of both eyes $\mathbf{e}^{\square} \in \mathbb{R}^4$.
    - The spherical harmonics illumination coefficients $\boldsymbol{\gamma}^{\square} \in \mathbb{R}^{27}$.
    - In total, there are $N_p = 261$ paremeters.

The Face Model

- The model is based on Blanz and Vetter's paper [1999].
- There's a template mesh with N vertices.
    - Let us denote the template mesh vertices by $\mathbf{a}_{\mathrm{geo}} \in \mathbb{R}^{3N}$.
- The model is then perturbed by two types of morphs.
    - The identity (i.e., face shape) morphs $\{\mathbf{b}_k^{\mathrm{geo}} \in \mathbb{R}^{3N} : k = 1, \dots, N_\alpha\}$.
    - The expression morphs $\{\mathbf{b}_k^{\mathrm{exp}} \in \mathbb{R}^{3N} : k = 1, \dots, N_\delta\}$.
- The morphed templated model is a function of $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$:

$$\mathbf{v}(\boldsymbol{\alpha}, \boldsymbol{\delta}) = \mathbf{a}_{\mathrm{geo}} + \sum_{i=1}^{k} \alpha_k \mathbf{b}_k^{\mathrm{geo}} + \sum_{i=1}^{k} \delta_k \mathbf{b}_k^{\mathrm{exp}}.$$

- Apperance of the mesh is determined by per-vertex diffuse albedo.
- There is a base albedo, denoted by $\mathbf{a}_{\mathrm{ref}} \in \mathbb{R}^{3N}$.
- The base albedo is perturbed by N_β change vectors $\{\mathbf{b}_k^{\mathrm{ref}} \in \mathbb{R}^3 N : k = 1, \dots, N_\beta\}$.
- The final per-vertex albedo is a function of $\boldsymbol{\beta}$:

$$\mathbf{r}(\boldsymbol{\beta}) = \mathbf{a}_{\mathrm{ref}} + \sum_{k=1}^{N_\beta} \beta_k \mathbf{b}_k^{\mathrm{ref}}.$$

- $\{\mathbf{b}_k^{\mathrm{geo}}\}$ and $\{\mathbf{b}_k^{\mathrm{ref}}\}$ are computed with PCA over 200 high-quality face scans.
- $\{\mathbf{b}_k^{\mathrm{exp}}\}$ are computed with PCA over the blendshapes of [Alexander et al. 2009] and [Cao et al. 2014b].
    - The blendshapes are transferred to the template mesh of Blanz and Vetter using deformation transfer [Sumner and Popovic et al. 2004].

Image Formation Model

- The 3D points $\mathbf{v}$ in object space is transformed into $\hat{\mathbf{v}}$ in camera space.

- It is then transformed into 2D points $\mathbf{p} = \Pi(\hat{\mathbf{v}})$ via a perspective projection.
- The perspective camera is assumed to be fixed for all frames.
- Consider the ith vertex of the deformed face mesh.
  - Let $\mathbf{r}_i$ denote the albedo of the vertex.
  - Let $\mathbf{n}_i$ denote the normal vector of the vertex.
  - The outgoing radiance from the vertex is given by:

    $$\mathbf{B}(\mathbf{r}_i, \mathbf{n}_i, \boldsymbol{\gamma}) = \mathbf{r}_i \otimes \sum_{b=1}^{B^2} \gamma_b Y_b(\mathbf{n}_i)$$

    where
    - $\otimes$ denotes component-wise multiplication,
    - $B = 3$ is the number bands of spherical harmonics to use, and
    - $Y_b : S^2 \to \mathbb{R}^3$ is the bth spherical harmonics basis function.

Face Reconstruction

- Let $\mathcal{X} = (R^{\square}, \mathbf{t}^{\square}, \boldsymbol{\alpha}^{\square}, \boldsymbol{\beta}^{\square}, \boldsymbol{\delta}^{\square}, \boldsymbol{\gamma}^{\square})$ denote a vector of the included parameters for the source/target video frame.
- Face reconstruction is done by minimizing the energy function

  $$E(\mathcal{X}) = w_{\text{photo}} E_{\text{photo}}(\mathcal{X}) + w_{\text{land}} E_{\text{land}}(\mathcal{X}) + w_{\text{reg}} E_{\text{reg}}(\mathcal{X})$$

  where:
  - $E_{\text{photo}}(\mathcal{X})$ takes into account photo-consistency between the rendering and the input frame.
    - They use the same loss function as in [Theis et al. 2016].
  - $E_{\text{land}}(\mathcal{X})$ takes into account landmark alignment.
    - The paper uses the tracker in [Saragih et al. 2011] to track 66 landmarks.
    - It's unclear how they define the loss function.
  - E_{reg} is a regularization term.
    - Enforces statistically plausible parameters based on the assumption of normally distributed data.
- $\mathcal{X}$ does not have the eye gaze parameter $\mathbf{e}^{\square}$. It is obtained directly from the landmark tracker.
- The identity of the face is estimated only for the first frame and kept constant afterwards while other parameters are recomputed every frame.
- Parameters are computed with the iteratively reweighted least square algorithm (IRLS) similar to the one used in Theis et al. [2016].

Synthetic Conditioning Input

- First, we relevant parameters (head pose, expression, eye gaze) from the source video to the target mesh.
- The images of the target mesh are rendered with hardware rasterization.

- To generate the fth frame of the output, the network takes a window of conditioning images $(\mathcal{C}_f, \mathcal{C}_{f-1}, \dots, \mathcal{C}_{f-10})$ of the 11 frames ending at the current frame as input.
- For each frame, we generate three different conditioning inputs:
  - a color rendering,
  - a correspondence image, and
  - an eye gaze image.
- The correspondence image simply encodes the index of the face model's vertices.
  - The paper textures the model with a gradient of colors based on the indices and just render it.
- The eye image has the region of the eye filled with white pixels.
  The locations of the pupils are rendered as blue circles on top of the white pixels.
- Taking all the images together, the input tensor has dimension $9N_w \times H \times W$ where $N_w = 11$ is the window size.

Rendering-to-Video Translation

- Input: the conditioning space-time video tensors as detailed in the last section.
- Output: a full frame of photo-realistic target video. This means it generates:
  - head motion,
  - facial expression,
  - eye gaze,
  - hair and body (hallucinated), and
  - background (halluciated).
- The network is trained from a specific target person and background.

Architecture

- It is a conditional generative adversarial network. It has two networks:
  - The space-time transformation network $\mathbf{T}$.
    - Takes in the $\mathbf{X}$, the $9N_w \times H \times W$ space-time tensor above.
    - Produces an image $\mathbf{T}(\mathbf{X})$ of the actor.
  - The discriminator $\mathbf{D}$.
    - Takes in $\mathbf{X}$ and an image which is supposed to be the real frame corresponding to the last synthetic frame in $\mathbf{X}$.
    - Produces a reality score for the image.
- The transformation network's architecture.
  - It is a U-Net.
  - It use cascade refinement [Chen and Koltun 2017].
  - Downsampling step:
    1. First, a 4×4 convolution with stride 2.
    2. Batch normalization.
    3. Leaky ReLU.

    - Upsampling step:
        1. A 4×4 convolution with upsampling factor for 2.
        2. Batch normalization.
        3. Drop out.
        4. ReLU.
        5. Then, two refinement steps with a 3×3 convolution followed by a a ReLU.
    - Last step is a $\tanh$ to produce the output in the $[-1, 1]$ range.
- The discriminator's architecture.
    - It extends the PatchGan architecture [Isola et al. 2017] to take into account the space-time volume.

Objective Functions

- The loss function for the translation network is given by

$$E_{\mathbf{X}}[\log(1 - \mathbf{D}(\mathbf{X}, \mathbf{T}(\mathbf{X})))] + \lambda E_{\mathbf{X},\mathbf{Y}}[\|\mathbf{X} - \mathbf{Y}\|_1].$$

  Here, $\mathbf{Y}$ is the ground truth image. The hyperparameter λ is set to 100.
    - Normally, for the original GAN loss, the adversarial term should be

$$-E_{\mathbf{X}}[\log(\mathbf{D}(\mathbf{X}, \mathbf{T}(\mathbf{X})))],$$

      which is the non-saturating varient of the term. The paper doesn't say it uses this varient.
    - I'm not so sure whether this is intentional or an editorial mistake.
- The loss function for the discriminator is given by

$$-E_{\mathbf{X},\mathbf{Y}}[\log \mathbf{D}(\mathbf{X}, \mathbf{Y})] - E_{\mathbf{X}}[\log(1 - \mathbf{D}(\mathbf{X}, \mathbf{D}(\mathbf{X})))].$$

Training

- The training corpus $\mathcal{T} = \{(\mathbf{X}_i, \mathbf{Y}_i) : 1 \leq i \leq f\}$ comes the video of the target person.
    - The ith frame becomes the ground truth image $\mathbf{Y}_i$.
    - The tensor $\mathbf{X}_i$ is generated by tracking the video frames and the generation of auxiliary outputs.
    - Typically, the authors use a video with two thousand frames $f = 2000$.
- Training settings.
    - Adam algorithm.
    - Learning rate of 2×10^{-4}.
    - $\beta_1 = 0.5$.
    - All other parameters have default value.
    - Bach size of 16
    - $31,000$ iterations (250 epochs).
- Weights are initialized by sampling from the normal distribution $\mathcal{N}(0, 0.2)$.