

网络结构：密集网络，残差网络，深层卷积网络

网络层：卷积，反卷积，反卷积之后再卷积 [deep video portraits]

L1 损失，可以描述生成后图片的相关性，保证 identity 一致性

cGAN 的损失，梯度惩罚等

帧间一致性损失，

写作方式：作比较，应用层的比较与技术层性能的比较。是否可量化为数值

参考论文： [1] Deep video portraits

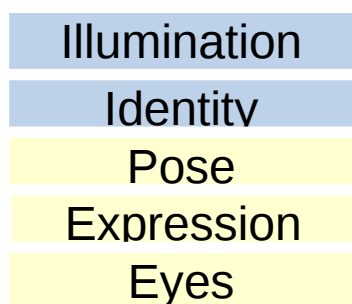
突出贡献： 空时编码结构作为时间一致的条件输入进行视频合成

表征人脸： face geometry+reflectance+motion+eye gaze/blinks

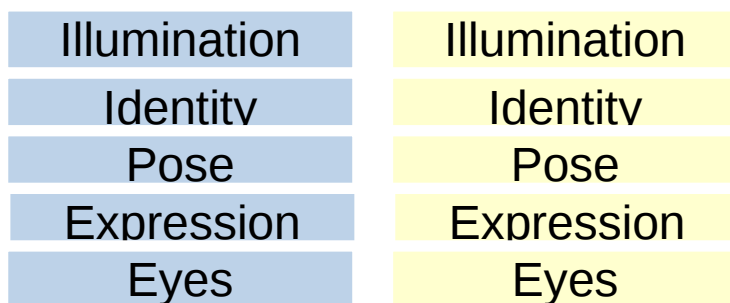
在视频中可以改善 view point variation 问题，利于二维投影

★ 3D monocular reconstruction

★ Target parameters representation

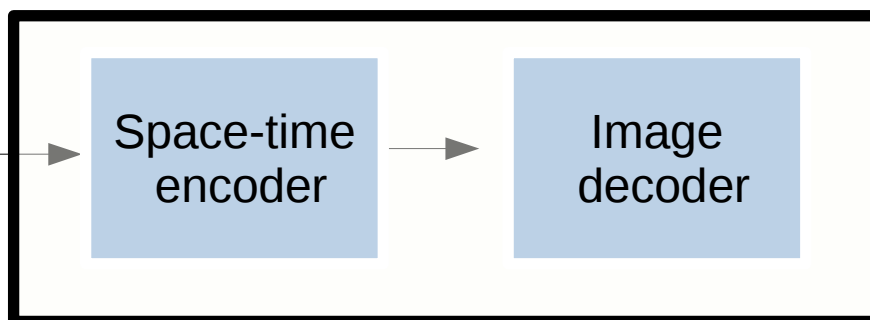


Modified Params



★ Conditional Generation

Conditional Input :
Shaded Colors +
Correspondence+Eyes



★ 关键步

pipeline

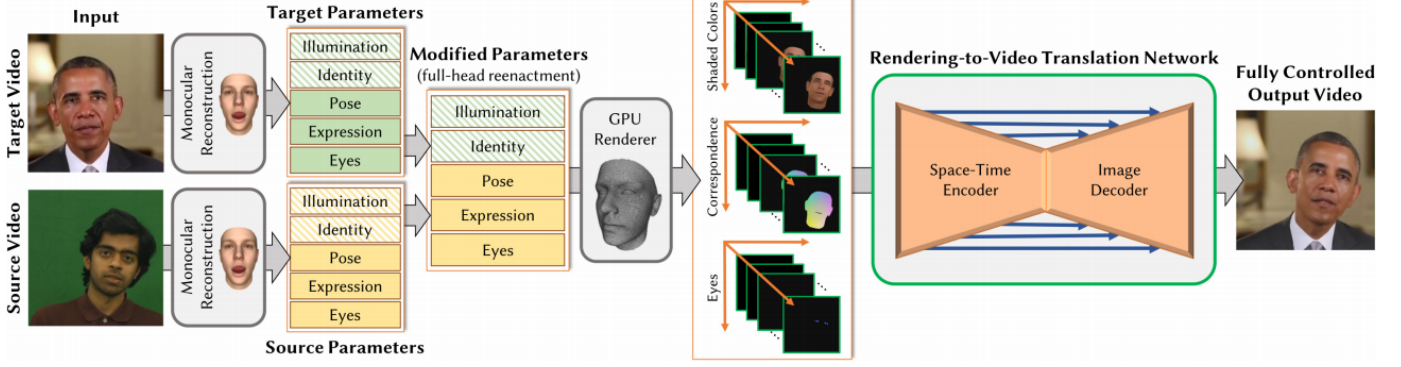


Fig. 2. Deep video portraits enable a source actor to fully control a target video portrait. First, a low-dimensional parametric representation (left) of both videos is obtained using monocular face reconstruction. The head pose, expression and eye gaze can now be transferred in parameter space (middle). We do not focus on the modification of the identity and scene illumination (hatched background), since we are interested in reenactment. Finally, we render conditioning input images that are converted to a photo-realistic video portrait of the target actor (right). *Obama* video courtesy of the White House (public domain).

Synthetic conditional input



Fig. 3. The synthetic input used for conditioning our rendering-to-video translation network: (1) colored face rendering under target illumination, (2) correspondence image, and (3) the eye gaze image.

Encoder-decoder architecs

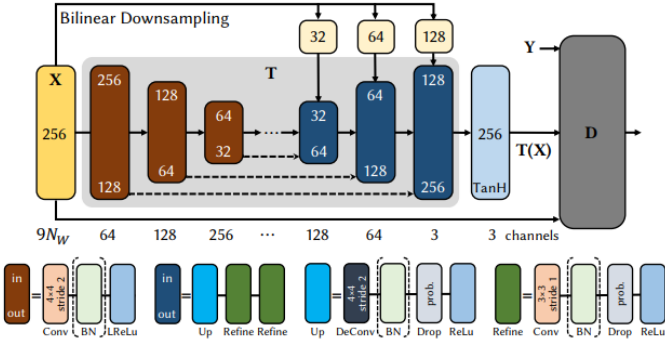


Fig. 4. Architecture of our rendering-to-video translation network for an input resolution of 256×256 : The encoder has 8 downsampling modules with (64, 128, 256, 512, 512, 512, 512, 512) output channels. The decoder has 8 upsampling modules with (512, 512, 512, 512, 256, 128, 64, 3) output channels. The upsampling modules use the following dropout probabilities (0.5, 0.5, 0.5, 0, 0, 0, 0, 0). The first downsampling and the last upsampling module do not employ batch normalization (BN). The final non-linearity (TanH) brings the output to the employed normalized $[-1, +1]$ -space.

Objective Function. We train in an adversarial manner to find the best rendering-to-video translation network:

$$T^* = \underset{T}{\operatorname{argmin}} \max_D E_{\text{cGAN}}(T, D) + \lambda E_{\ell_1}(T). \quad (5)$$

This objective function comprises an adversarial loss $E_{\text{cGAN}}(T, D)$ and an ℓ_1 -norm reproduction loss $E_{\ell_1}(T)$. The constant weight of $\lambda = 100$ balances the contribution of these two terms. The adversarial loss has the following form:

$$E_{\text{GAN}}(T, D) = \mathbb{E}_{X, Y} [\log D(X, Y)] + \mathbb{E}_X [\log (1 - D(X, T(X)))]. \quad (6)$$

fooling the discriminator. The ℓ_1 -norm loss penalizes the distance between the synthesized image $T(X)$ and the ground-truth image Y , which encourages the sharpness of the synthesized output:

$$E_{\ell_1}(T) = \mathbb{E}_{X, Y} [\|Y - T(X)\|_1]. \quad (7)$$

