VAE/WGAN-BASED IMAGE REPRESENTATION LEARNING FOR POSE-PRESERVING SEAMLESS IDENTITY REPLACEMENT IN FACIAL IMAGES

BU ECE

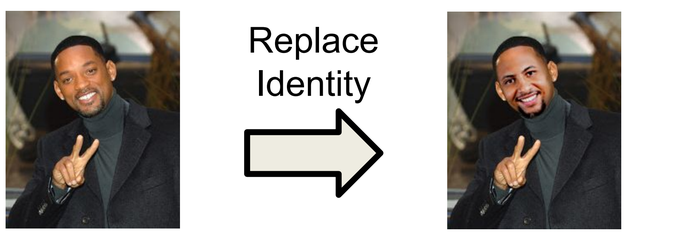
Hiroki Kawai, Jiawei Chen, Janusz Konrad, and Prakash Ishwar
{hirokik, garychen, jkonrad, pi} @bu.edu
Department of Electrical and Computer Engineering, Boston Univeristy, Boston, MA, USA

LESA

NVIDIA

Motivation

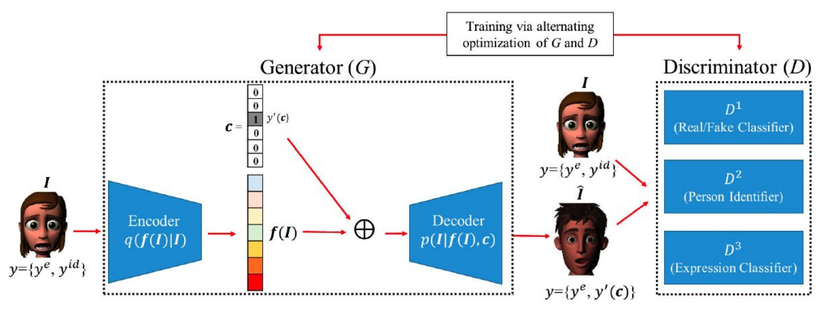
- **Smartroom of the future**: could improve energy efficiency, health outcomes and productivity by recognizing activities of occupants
- **Standard approach**: video cameras
- **Problem**: privacy concerns
- **Proposed solution**: seamlessly replace occupant's appearance while preserving other useful information like expression, pose, etc.



- **State of the art:** PPRL-VGAN [1] deep neural network for identity replacement that preserves the **facial expression** of an individual

- **Contributions:**
  - PPRL-VGAN framework to preserve **headpose**
  - **Inception modules** to improve image quality
  - **WGAN** + modified cost function (image reconstruction cost) to improve training stability and image quality

PPRL-VGAN



Architecture of PPRL-VGAN [1]

Generator: based on Variational Autoencoder:
- encoder converts input image into a latent vector representation
- decoder synthesizes a new realistic-looking image with specified identity from a latent vector

Discriminator: 3 prediction objectives

D^1 - Is image **real or fake**?

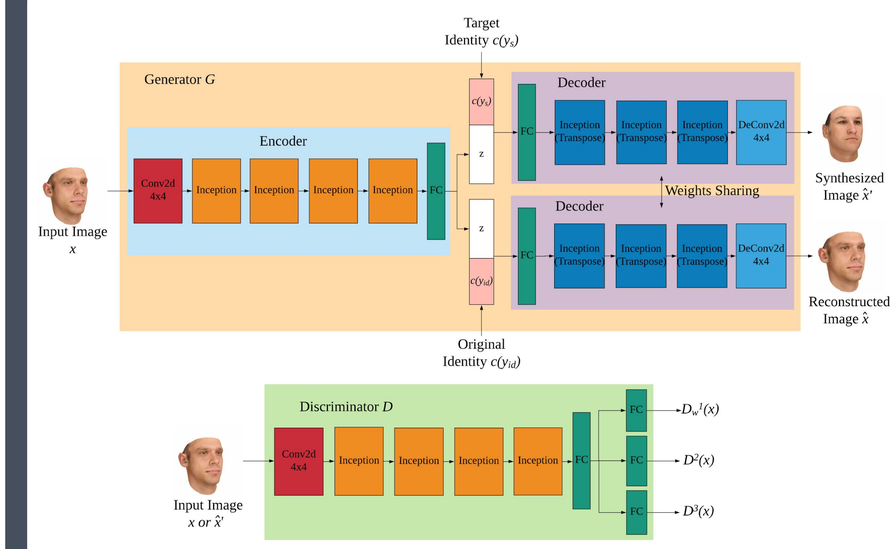
D^2 - **Identity**

D^3 - **Facial expression**

[1] J. Chen, J. Konrad, and P. Ishwar, "VGAN-based image representation learning for privacy-preserving facial expression recognition," *CVPR COPS Workshp* 2018.

Proposed Methodology for Head Pose Estimation

(1) PPRL-VGAN for headpose estimation:



(2) **Inception modules:** contain **3 convolutional-layer branches** with different filter sizes; branch outputs are concatenated

(3) Improved training method:

- **Wasserstein GAN (WGAN)**: leverages Earth-Mover distance (instead of Jensen-Shannon divergence) via **gradient penalty** in discriminator loss
- **Image reconstruction cost**: compares input image and generated image to improve image quality
- ***Generator Loss***: encourages synthesis of realistic images with new identity and original headpose

$$L_G = E\big[-D_\omega^1(G(\mathbf{x},\mathbf{c}(y_s)))\big] + E\big[-\log(D_{y_s}^2(G(\mathbf{x},\mathbf{c}(y_s))))\big] + E\Big[\sum_{i=1}^{3}|y_{pose}^i - D_i^3(G(\mathbf{x},\mathbf{c}(y_s)))|\Big] + E\big[||G(\mathbf{x},\mathbf{c}(y_{id}))-\mathbf{x}||_2^2\big] + D_{KL}(q(\mathbf{z}|\mathbf{x})||r(\mathbf{z}))$$

- ***Discriminator Loss***: encourages accurate prediction of identity, expresion and real vs synthetic detection

$$L_D = E\big[-D_\omega^1(\mathbf{x}) + D_\omega^1(G(\mathbf{x},\mathbf{c}(y_s)))\big] + E\big[-\log(D_{y_{id}}^2(\mathbf{x}))\big] + E\Big[\sum_{i=1}^{3}|y_{pose}^i - D_i^3(\mathbf{x})|\Big] + E\big[(||\nabla_{\mathbf{x}} D_\omega^1(\mathbf{x})||_2 - 1)^2\big]$$

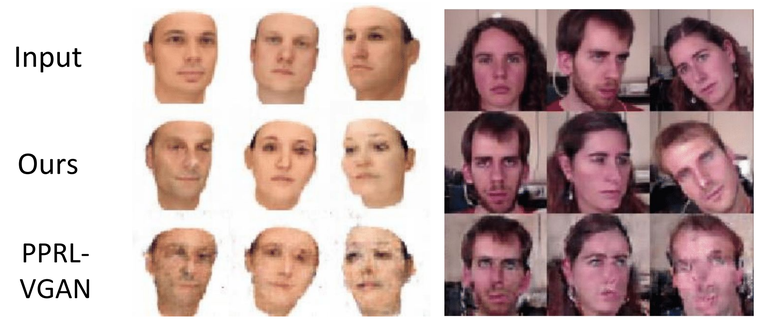
:WGAN cost :Identity cost :Head-pose cost

:Image Reconstruction :Regularization :Gradient penalty

- Training alternates between minimizing L_G and maximizing L_D
- These loss functions are minimized *via* **Adam optimization**

Experimental Results

Qualitative evaluation: UPNA Head Pose Database (Cropped images of centered faces resized to 64x64 pixels)



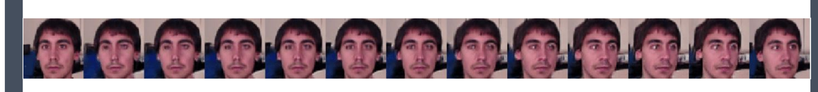
Quantitative evaluation: Privacy protection evaluated by training another neural network to predict identity under 3 attack scenarios :

Attack Scenario	Identification (%)		Headpose MAE (°)	
Method	Ours	PPRL-VGAN	Ours	PPRL-VGAN
Privacy Unconstrained	99.97		0.69	
Training: Original Dataset Test: Synthesized Images	10.23	9.92	2.25	3.57
Training: Synthesized Images Test: Synthesized Images	23.31	21.64	1.81	2.90
Training: Latent Vectors Test: Latent Vectors	21.33	23.71	2.21	2.76

Identity/head-pose morphing: The generative ability of our model is evaluated by identity and head-pose morphing:

Identity morphing:

Head-pose morphing:



Conclusions

- Our method synthesizes realistic face images with a desired identity and improved image quality compared to a state-of-the-art method.
- We achieve performance competitive with a state-of-the-art method for learning an identity-invariant image representation.
- Our model can be applied to other image tasks such as pose or face morphing.

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research and support from the BU Undergraduate Research Opportunities Program