Few-Shot Adversarial Learning of Realistic Neural Talking Head Models

Andrey Gusev

Higher School of Economics

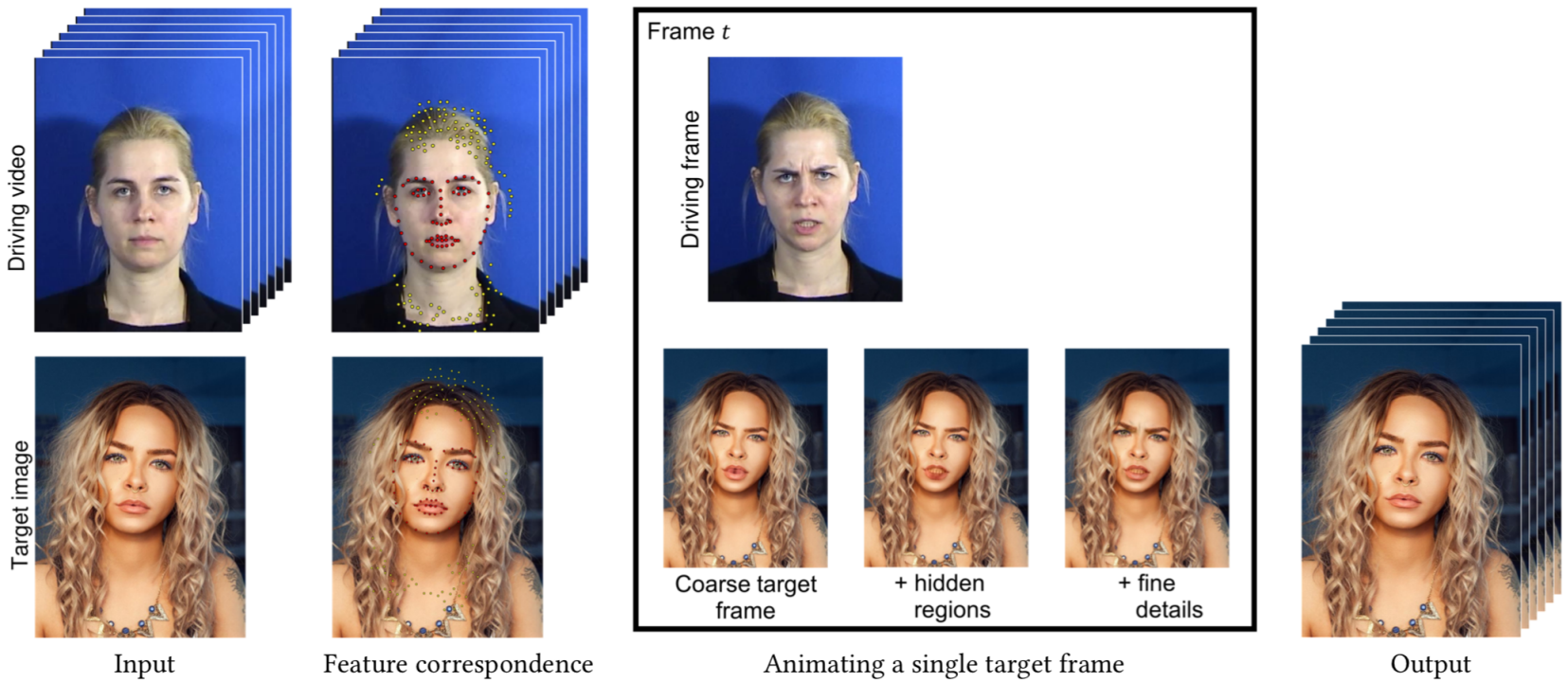
aagusev_2@edu.hse.ru

April 3, 2020

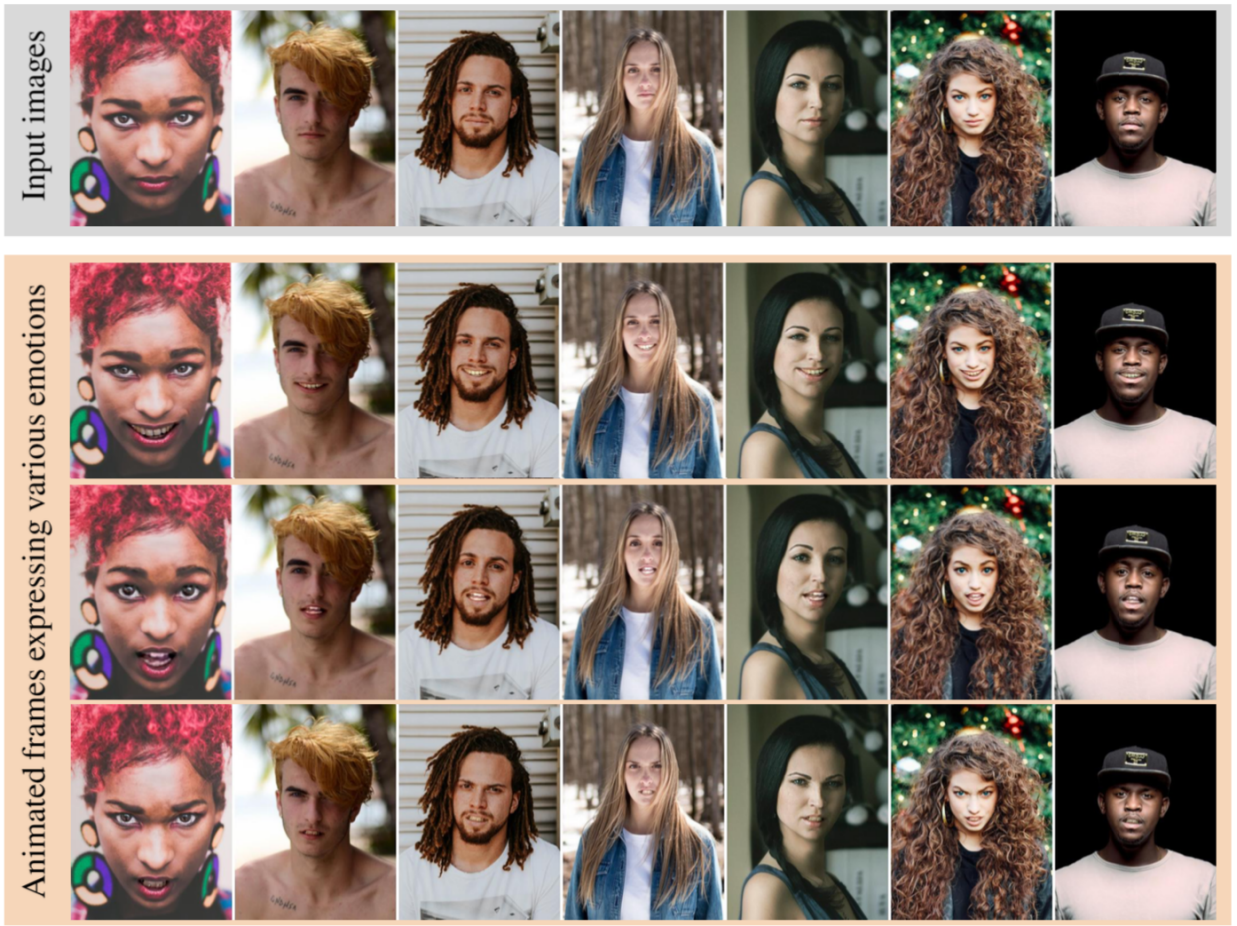
Overview

Problem formulation

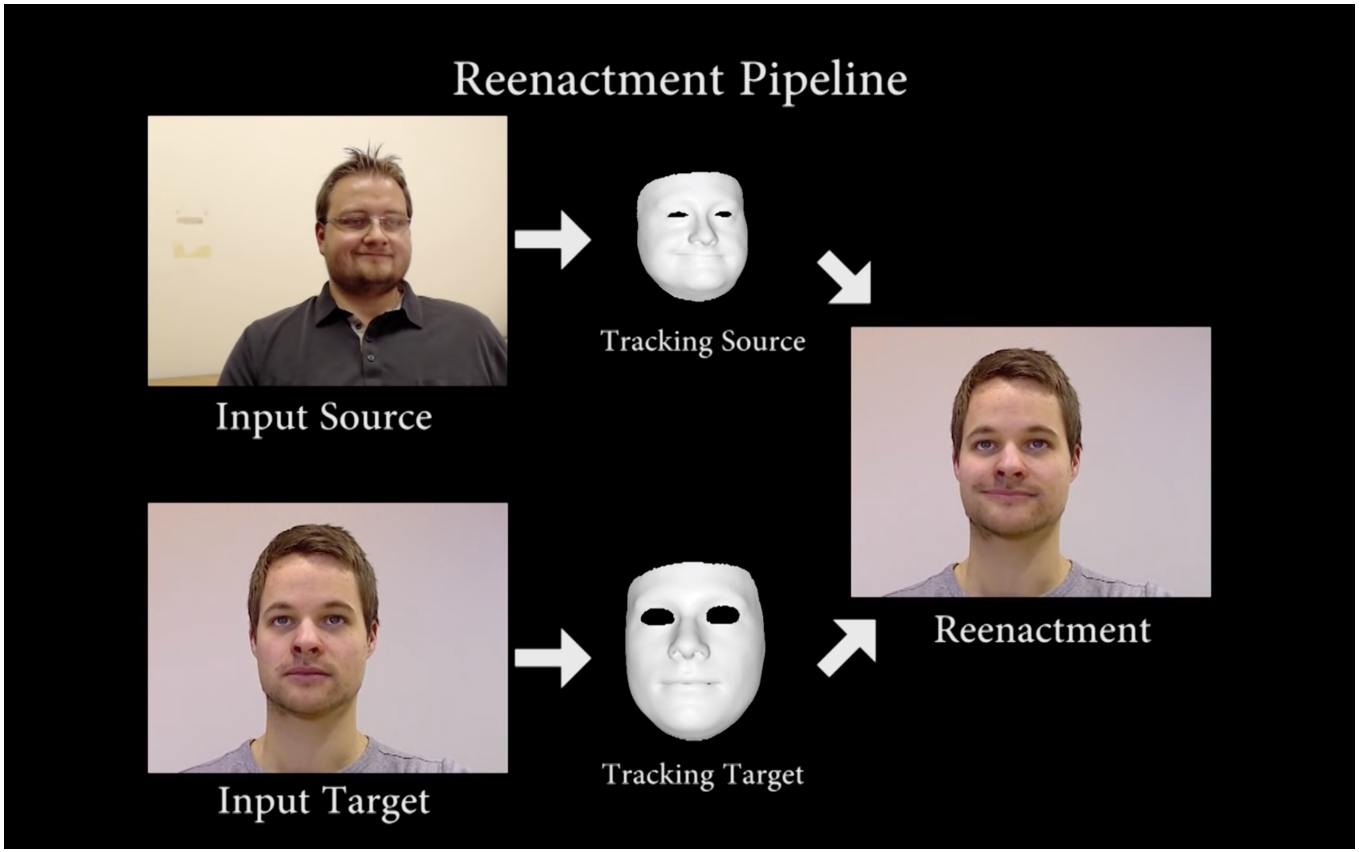
Warping method



Warping examples



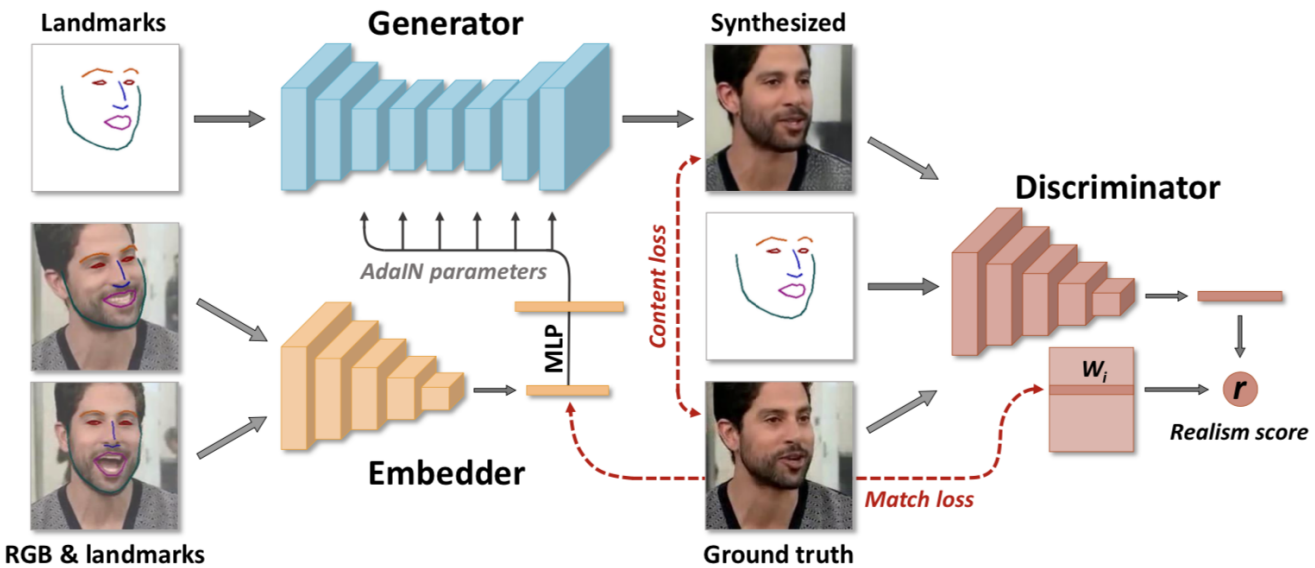
Face2Face (2016)



Stage purpose

During the meta-learning stage of authors approach, the parameters of all three networks are trained in an adversarial fashion. It is done by simulating episodes of K-shot learning (K = 8 in original experiments).

General structure



Embedder

$$E(x_i(s), y_i(s); \phi) \to \hat{e}_i(s)$$

$x_i(s)$ – a video frame
$y_i(s)$ – an associated landmark image
ϕ – network parameters that are learned in the meta-learning stage
$\hat{e}_i(s)$ – an N-dimensional vector

Generator

$$G(y_i(t), \hat{e}_i; \psi, \mathrm{P}) \to \hat{x}_i(t)$$

$y_i(t)$ – an associated landmark image
$\hat{e}_i$ – the predicted video embedding

$$\hat{e}_i = \frac{1}{K} \sum_{k=1}^{K} E(x_i(s_k), y_i(s_k); \phi)$$

ψ – the person-generic parameters
$\hat{\psi}_i$ – the person-specific parameters
$\mathrm{P} : \hat{\psi}_i = \mathrm{P}\hat{e}_i$
$\hat{x}_i(t)$ – a synthesized video frame

Discriminator

$$D(x_i(t), y_i(t), i; \theta, \mathrm{W}, \mathrm{w}_0, b) \to r$$

$x_i(t)$ – a video frame
$y_i(t)$ – an associated landmark image
i – the index of the training sequence

The discriminator contains a ConvNet part $V(x_i(t), y_i(t); \theta)$

r – a single scalar realism score, based on the output of its ConvNet and the parameters $\mathrm{W}, \mathrm{w}_0, b$

Embedder and generator objective

$$\mathcal{L}(\phi, \psi, \mathsf{P}, \theta, \mathsf{W}, \mathsf{w}_0, b) = \mathcal{L}_{\text{CNT}}(\phi, \psi, \mathsf{P}) + \mathcal{L}_{\text{ADV}}(\phi, \psi, \mathsf{P}, \theta, \mathsf{W}, \mathsf{w}_0, b) + \mathcal{L}_{\text{MCH}}(\phi, \mathsf{W})$$

$\mathcal{L}_{\text{CNT}}$ – the content term
$\mathcal{L}_{\text{ADV}}$ – the adversarial term
$\mathcal{L}_{\text{MCH}}$ – the embedding match term

Content term

It measures the distance between the ground truth image $x_i(t)$ and the reconstruction $\hat{x}_i(t)$ using the perceptual similarity measure. Such kind of loss is based on high-level features extracted from pretrained networks.

Adversarial term

$$\mathcal{L}_{\text{ADV}}(\phi, \psi, \mathsf{P}, \theta, \mathsf{W}, \mathsf{w}_0, b) = -r + \mathcal{L}_{\text{FM}} = \\ - D(\hat{x}_i(t), y_i(t), i; \theta, \mathsf{W}, \mathsf{w}_0, b) + \mathcal{L}_{\text{FM}}$$

$$D(\hat{x}_i(t), y_i(t), i; \theta, \mathsf{W}, \mathsf{w}_0, b) = V(\hat{x}_i(t), y_i(t), i; \theta)^T (\mathsf{W}_i + \mathsf{w}_0) + b$$

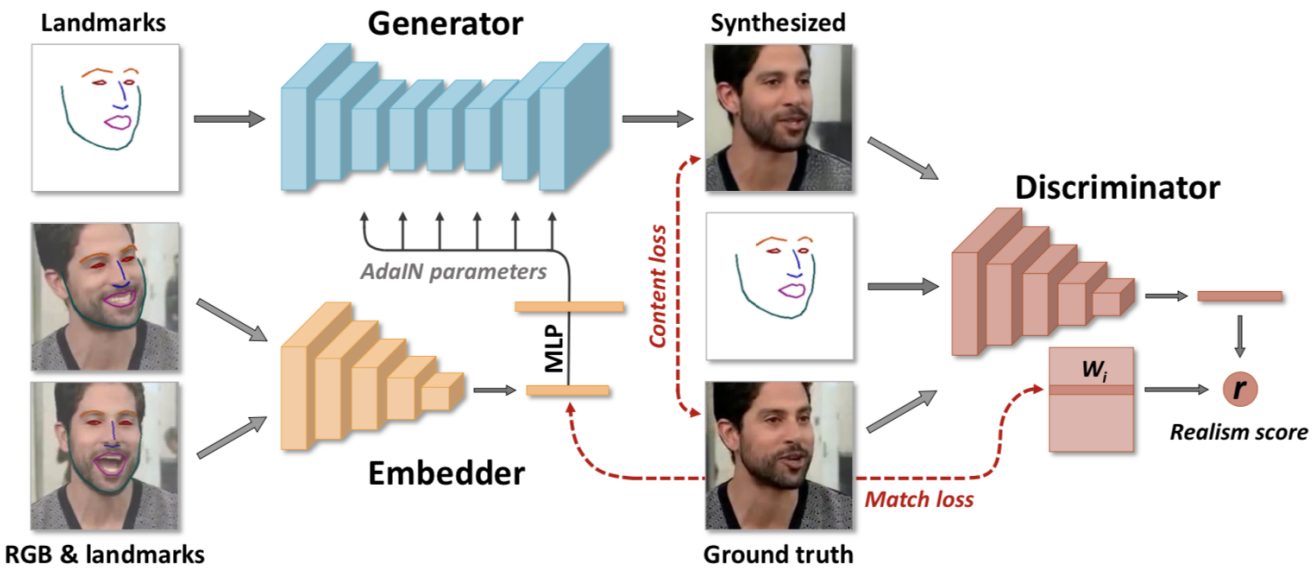
Match term

It encourages the similarity of the two types of embeddings by penalizing the L_1-difference between $E(x_i(s_k), y_i(s_k); \phi)$ and W_i.

Discriminator objective

$$\mathcal{L}_{\text{DSC}}(\phi, \psi, \mathsf{P}, \theta, \mathsf{W}, \mathsf{w}_0, b) =$$
$$+ \max(0, 1 + D(\hat{x}_i(t), y_i(t), i; \phi, \psi, \theta, \mathsf{W}, \mathsf{w}_0, b))$$
$$+ \max(0, 1 - D(x_i(t), y_i(t), i; \theta, \mathsf{W}, \mathsf{w}_0, b))$$

General structure



Stage purpose

The system is learned in a few-shot way, assuming that T training images $x_1, x_2, \ldots, x_T$ (e.g. T frames of the same video) are given and that $y_1, y_2, \ldots, y_T$ are the corresponding landmark images.

Replaced components

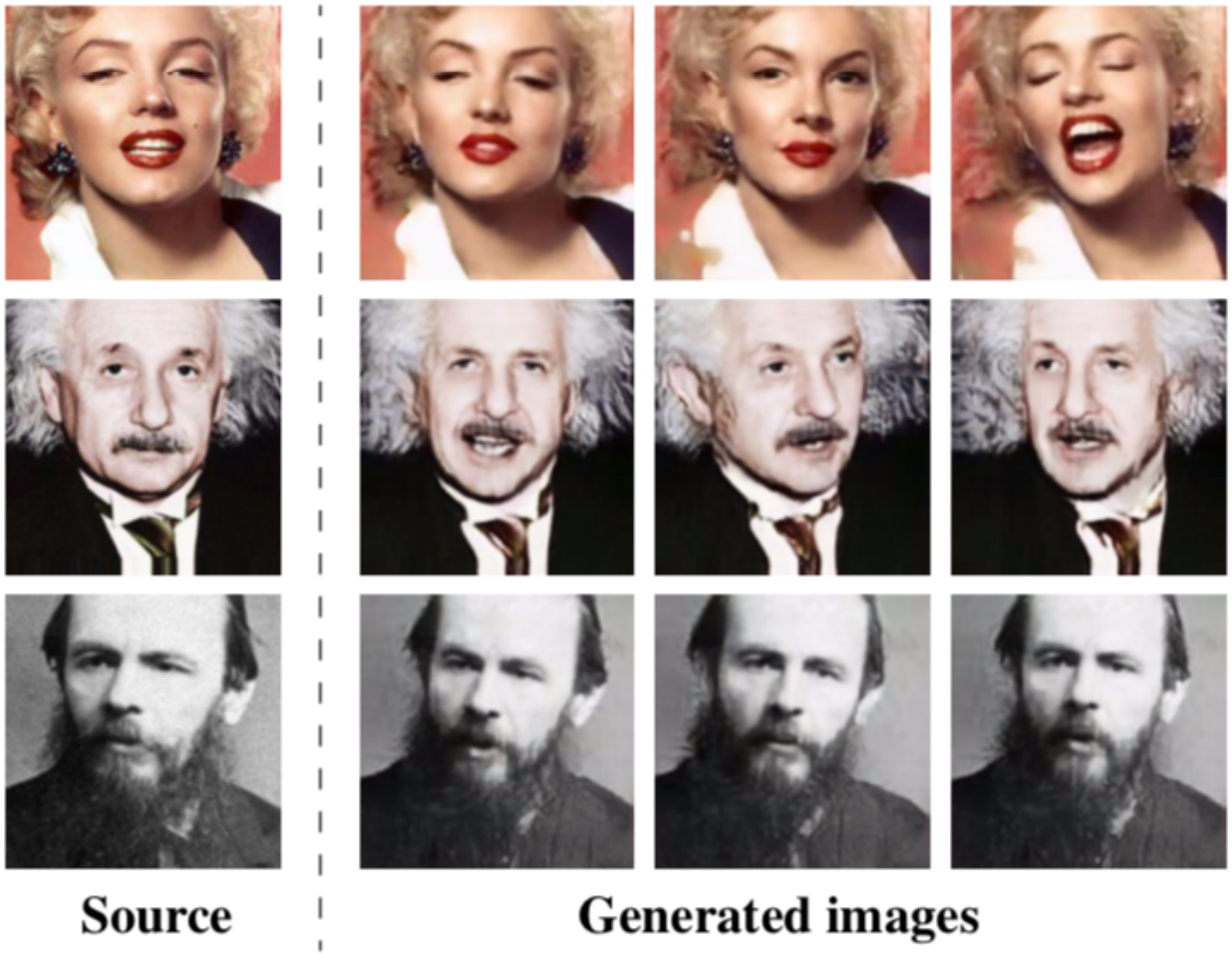
- Generator: $G'(y(t); \phi, \psi')$, where $\psi' = \mathsf{P}\hat{e}_{\mathrm{NEW}}$
- Discriminator: $D'(x(t), y(t); \theta, \mathsf{w}', b)$, where $\mathsf{w}' = \mathsf{w}_0 + \hat{e}_{\mathrm{NEW}}$

$$D'(\hat{x}(t), y(t); \theta, \mathsf{w}', b) = V(\hat{x}(t), y(t), \theta)^T \mathsf{w}' + b$$

Method (T)	FID↓	SSIM↑	CSIM↑	USER↓
VoxCeleb1				
X2Face (1)	45.8	**0.68**	**0.16**	0.82
Pix2pixHD (1)	**42.7**	0.56	0.09	0.82
Ours (1)	43.0	0.67	0.15	**0.62**
X2Face (8)	51.5	**0.73**	**0.17**	0.83
Pix2pixHD (8)	**35.1**	0.64	0.12	0.79
Ours (8)	38.0	0.71	**0.17**	**0.62**
X2Face (32)	56.5	**0.75**	0.18	0.85
Pix2pixHD (32)	**24.0**	0.70	0.16	0.71
Ours (32)	29.5	0.74	**0.19**	**0.61**
VoxCeleb2				
Ours-FF (1)	**46.1**	0.61	**0.42**	**0.43**
Ours-FT (1)	48.5	**0.64**	0.35	0.46
Ours-FF (8)	42.2	0.64	**0.47**	0.40
Ours-FT (8)	**42.2**	**0.68**	0.42	**0.39**
Ours-FF (32)	40.4	0.65	**0.48**	0.38
Ours-FT (32)	**30.6**	**0.72**	0.45	**0.33**


1
8
32
T
Source
Ground truth
X2Face
Pix2pixHD
Ours


1
8
32
T
Source
Ground truth
Ours-FF
Ours-FT
before fine-tuning
Ours-FT
after fine-tuning


Source
Generated images

References

Zakharov et al. (2019)

Few-Shot Adversarial Learning of Realistic Neural Talking Head Models

Averbuch et al. (2017)

Bringing portraits to life

Isola et al. (2016)

Image-to-Image Translation with Conditional Adversarial Networks

Questions

1. What is the purpose of the Embedder network in the considered model?
2. What does the Generator expect as an input and what does it output?
3. Describe general structure of the model, three main networks.