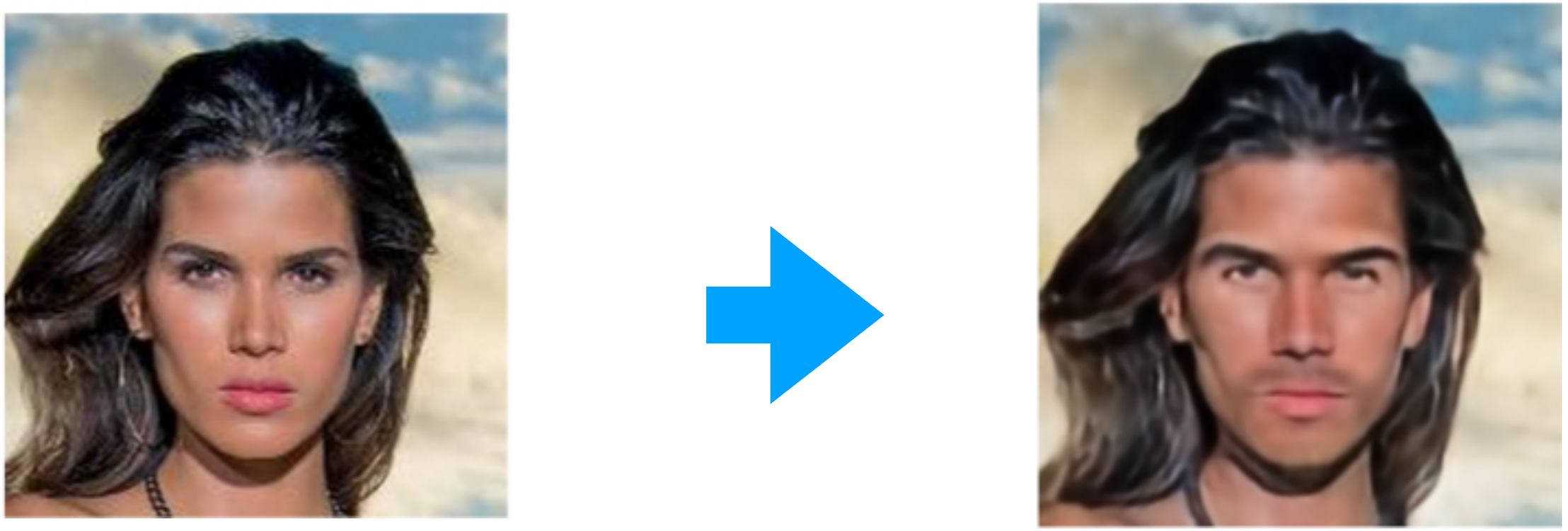
Manipulating Images by Sliding Attributes

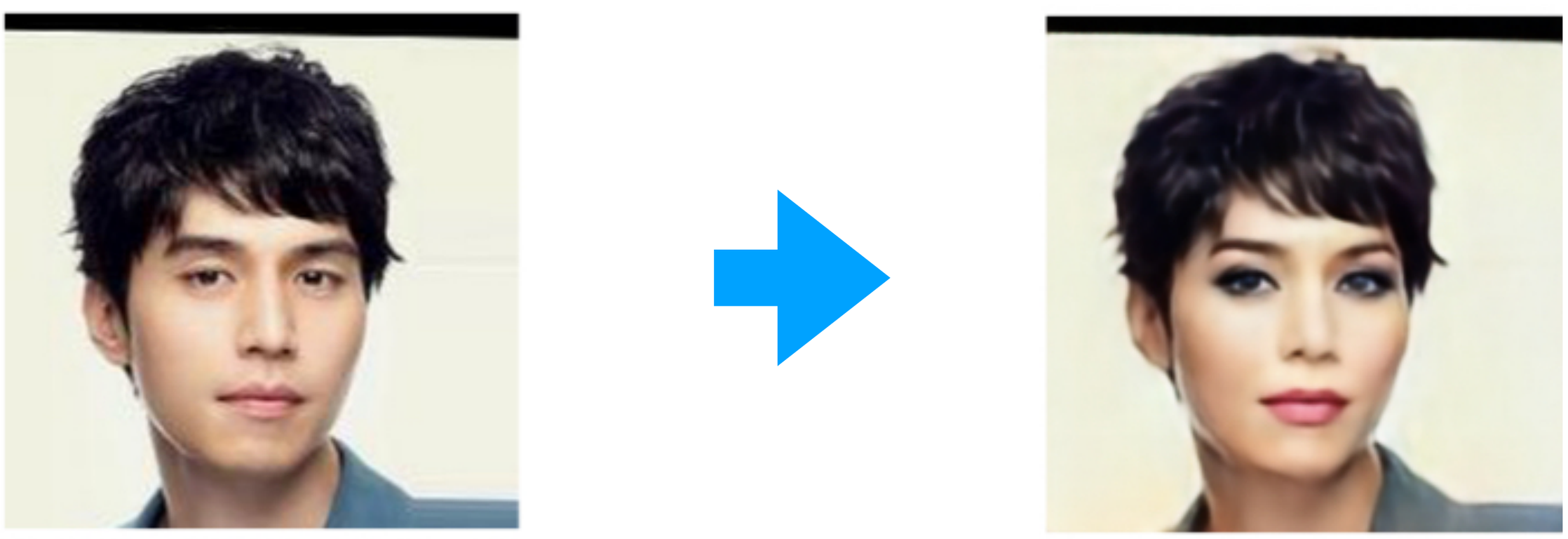
Павел Латышев 151

Цель работы

Менять параметры фото лица:

- пол
- возраст
- выражение лица
- наличие очков





В чем сложность?

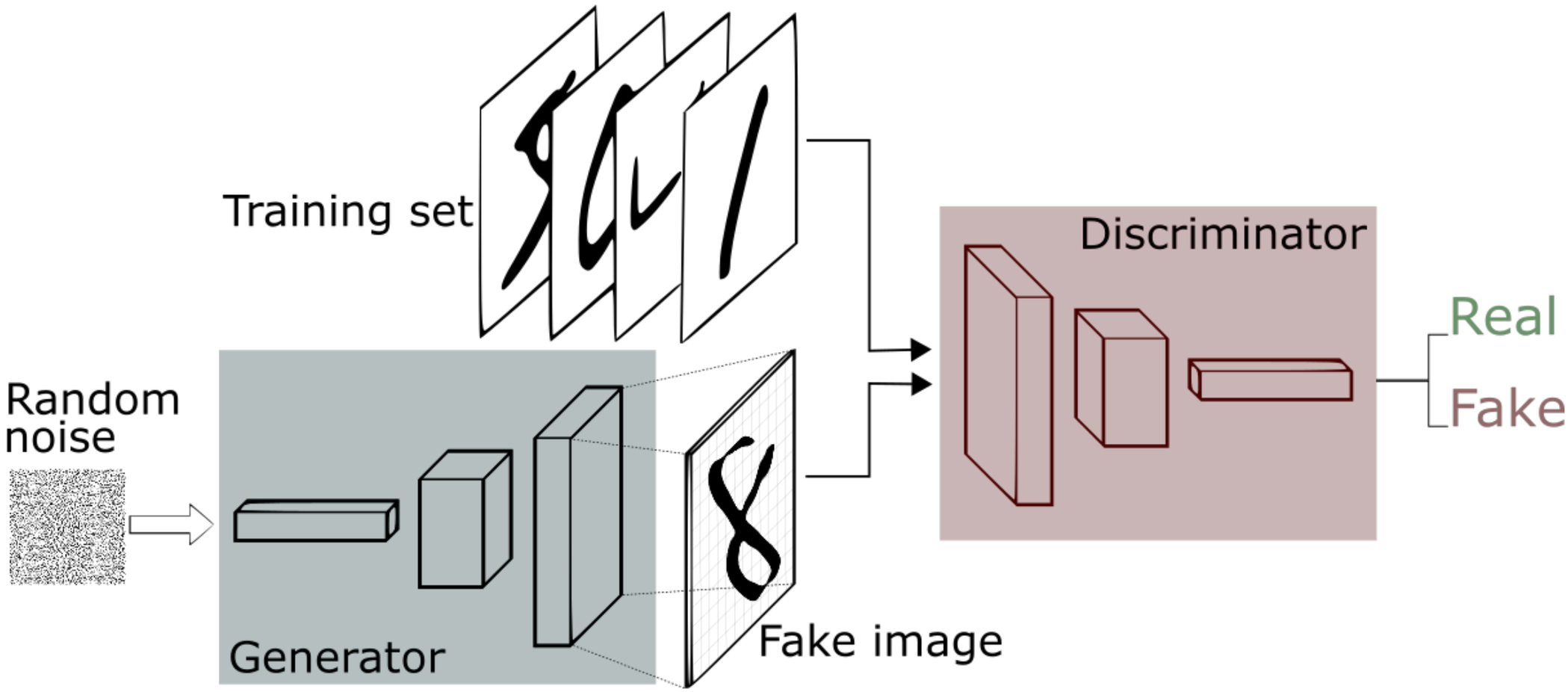
- Дорого делать обучающую выборку для разных выражений лица, наличия очков, возраста

- Невозможно собрать выборку для пола

Какие данные есть?

Выборка - фото лиц с размеченными параметрами

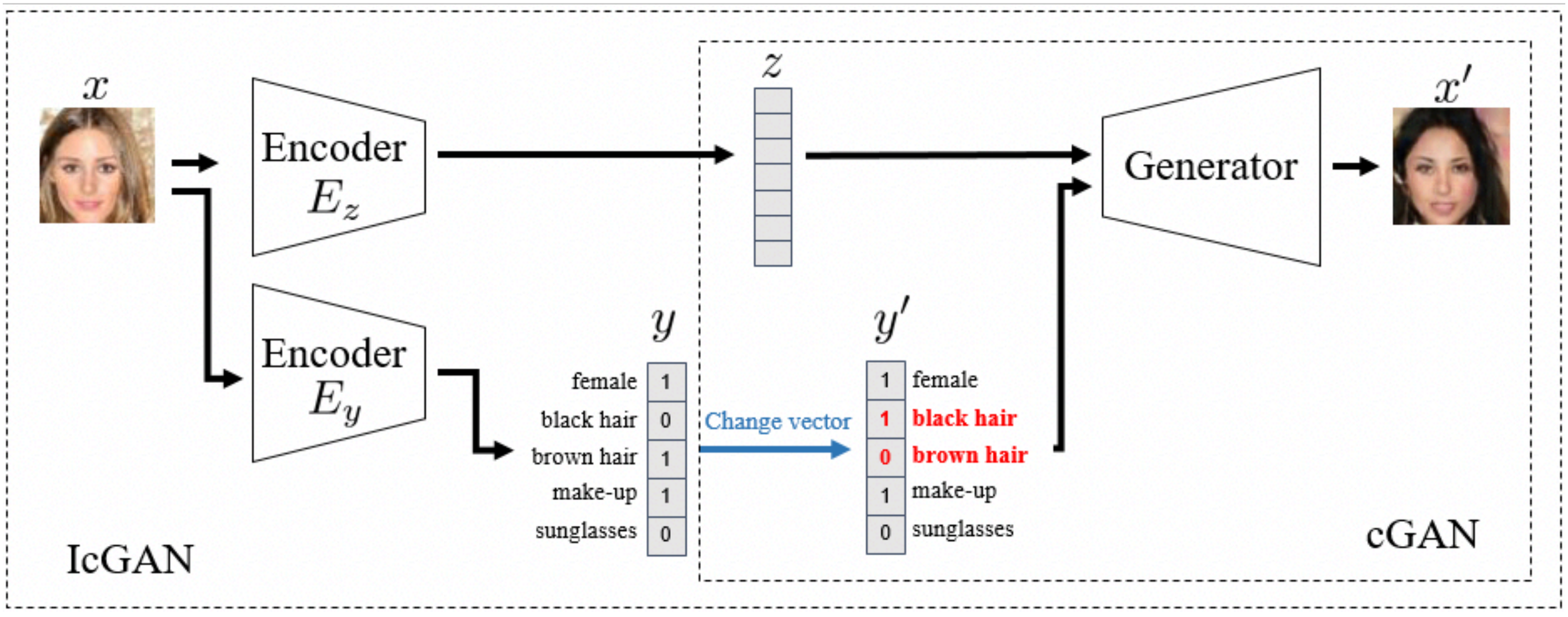
GAN



$$\min_g \max_d v(\theta_g, \theta_d) = \mathbb{E}_{x \sim p_{data}}[\log D(x)] + \mathbb{E}_{z \sim p_z}[\log(1 - D(G(z)))]$$

$$\min_g \max_d v(\theta_g, \theta_d) = \mathbb{E}_{x,y \sim p_{data}}[\log D(x,y)] + \mathbb{E}_{z \sim p_z, y' \sim p_y}[\log(1 - D(G(z,y'),y'))]$$

Invertible Conditional GAN



$$\mathcal{L}_{ez} = \mathbb{E}_{z\sim p_z, y'\sim p_y} \|z - E_z(G(z, y'))\|_2^2$$

$$\mathcal{L}_{ey} = \mathbb{E}_{x,y\sim p_{data}} \|y - E_y(x)\|_2^2$$

Энкодер

- SNG - один энкодер с двумя выходами
- IND - два независимых энкодера
- IND-COND - два декодера, E_z зависит от выхода E_y

Table 2: Encoder IND architecture. Last two layers have different sizes depending on the encoder (z for E_z or y for E_y). n_y represents the size of y.

Operation	**Kernel**	**Stride**	**Filters**	**BN**	**Activation**
Convolution	5×5	2×2	32	Yes	ReLU
Convolution	5×5	2×2	64	Yes	ReLU
Convolution	5×5	2×2	128	Yes	ReLU
Convolution	5×5	2×2	256	Yes	ReLU
Fully connected	-	-	z: 4096, y: 512	Yes	ReLU
Fully connected	-	-	z: 100, y: n_y	No	None

cGAN

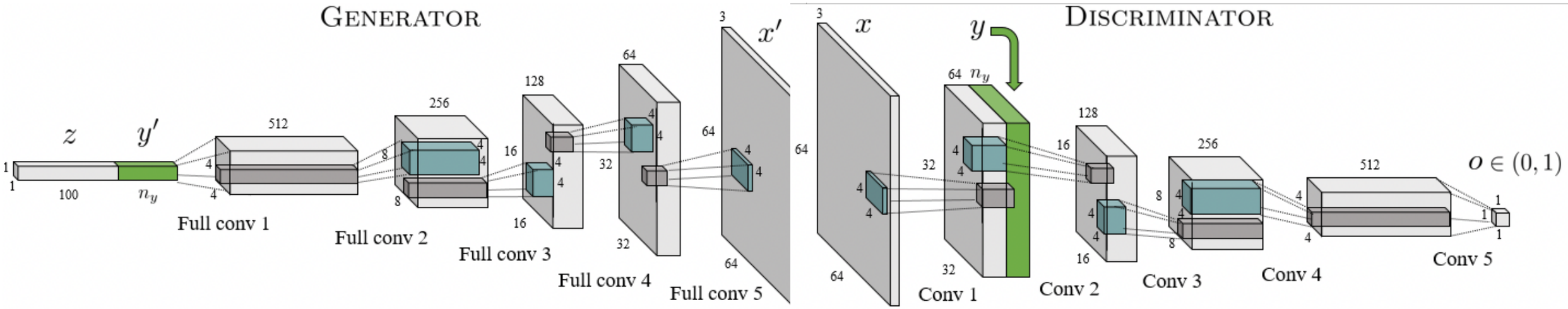


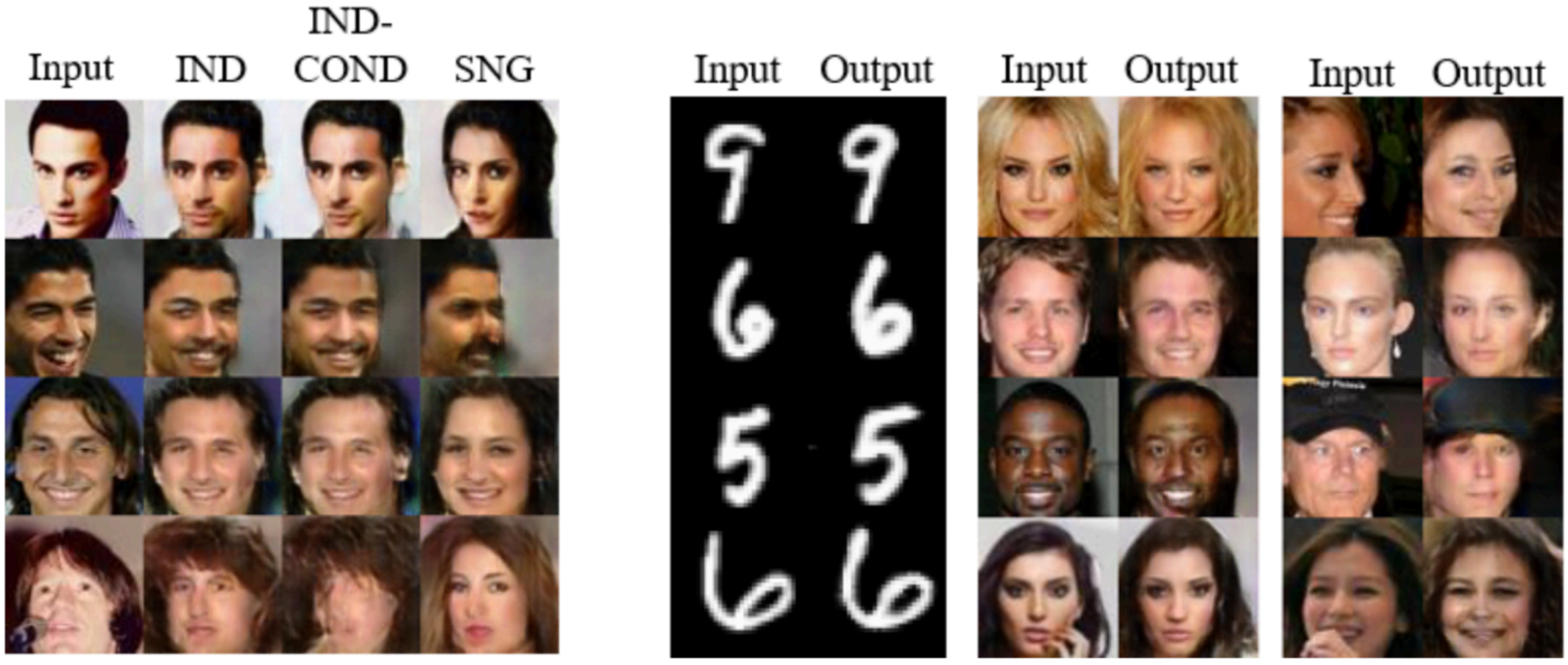
Table 1: Detailed generator and discriminator architecture

Generator						Discriminator					
Operation	**Kernel**	**Stride**	**Filters**	**BN**	**Activation**	**Operation**	**Kernel**	**Stride**	**Filters**	**BN**	**Activation**
Concatenation	*Concatenate z and y' on 1st dimension*					Convolution	4×4	2×2	64	No	Leaky ReLU
Full convolution	4×4	2×2	512	Yes	ReLU	Concatenation	*Replicate y and concatenate to 1st conv. layer*				
Full convolution	4×4	2×2	256	Yes	ReLU	Convolution	4×4	2×2	128	Yes	Leaky ReLU
Full convolution	4×4	2×2	128	Yes	ReLU	Convolution	4×4	2×2	256	Yes	Leaky ReLU
Full convolution	4×4	2×2	64	Yes	ReLU	Convolution	4×4	2×2	512	Yes	Leaky ReLU
Full convolution	4×4	2×2	3	No	Tanh	Convolution	4×4	1×1	1	No	Sigmoid

Обучение модели

1. Обучается cGAN
2. Создается датасет сгенерированных изображений
3. Обучается Ez выдавать внутреннее представление z для сгенерированных картинок
4. Обучается Ey выдавать вектор атрибутов

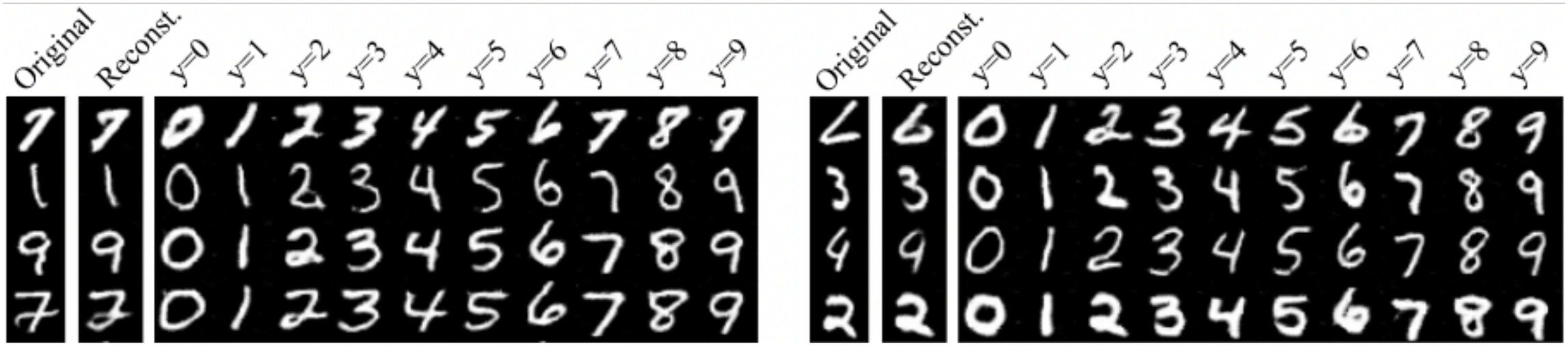
Сравнение энкодеров



(a) (b)

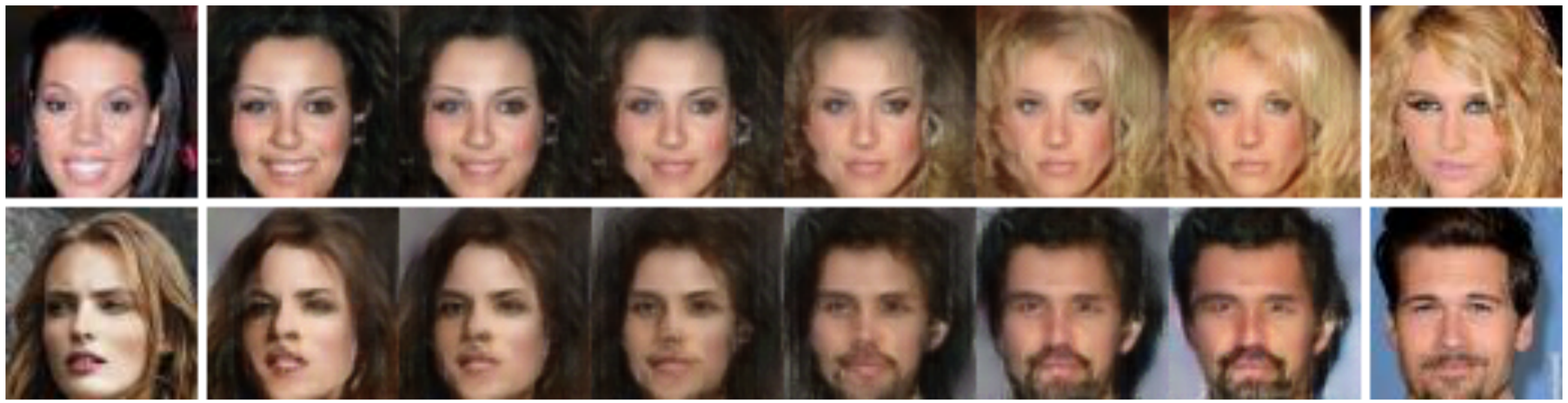
Figure 4: (a) Comparison of different encoder configurations, where IND yields the most faithful reconstructions. (b) Reconstructed samples from MNIST and CelebA using IND configuration.

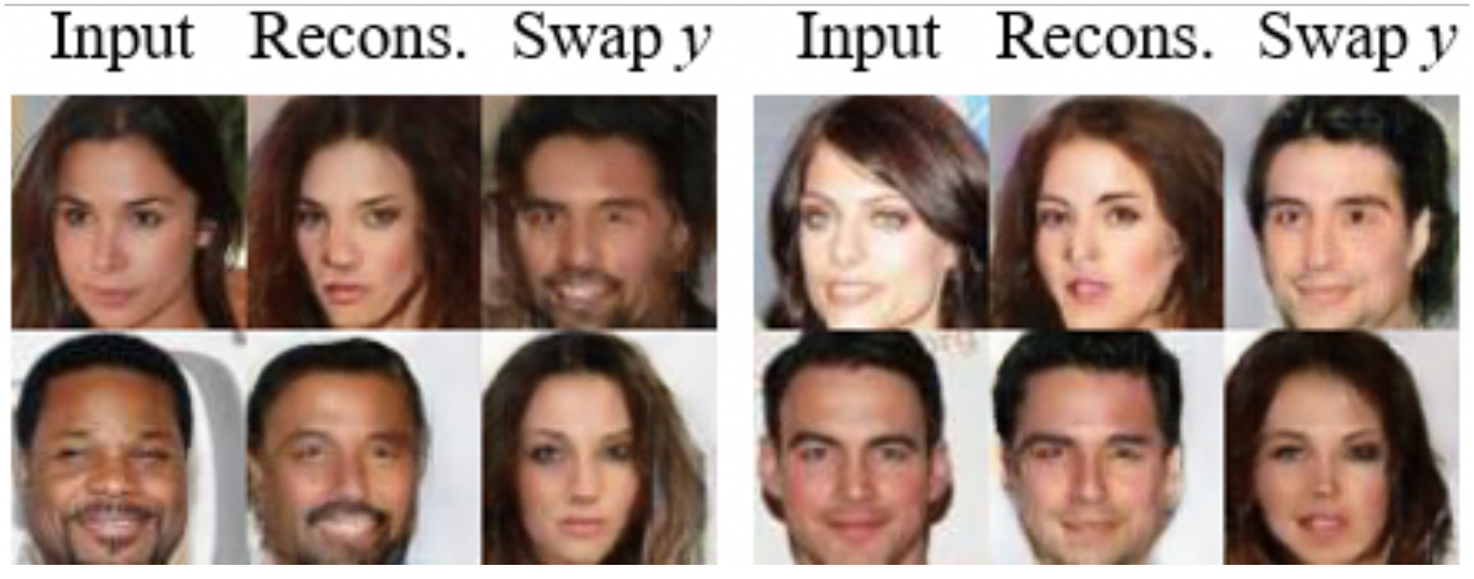
Результаты


Original
Reconst.
y=0
y=1
y=2
y=3
y=4
y=5
y=6
y=7
y=8
y=9
Original
Reconst.
y=0
y=1
y=2
y=3
y=4
y=5
y=6
y=7
y=8
y=9


Original
Reconstruction
Bald
Bangs
Black hair
Blonde
Eyeglasses
Heavy makeup
Gender change
Pale skin
Smiling

Результаты





Fader Network

Fader netword

X **- пространство изображений**

$Y = \{0,1\}^n$ **- пространство параметров**

$D = \{(x^1, y^1), \ldots, (x^m, y^m)\}^n$ **- обучающая выборка**

Encoder-decoder

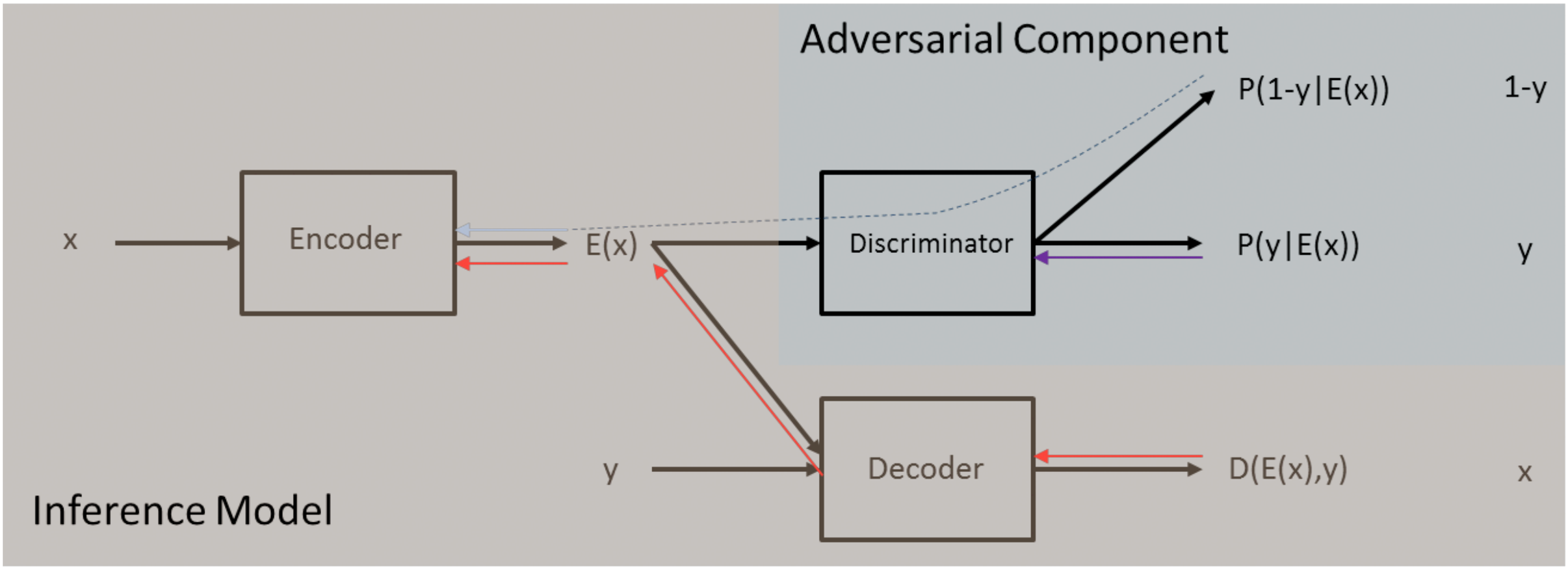
$E_{\theta_{enc}} : X \to \mathbb{R}^N$**- энкодер**

$D_{\theta_{dec}} : (\mathbb{R}^N, Y) \to X$**- декодер**

$$\mathscr{L}_{AE}(\theta_{enc}, \theta_{dec}) = \frac{1}{m} \sum_{(x,y)\in D} ||D_{\theta_{dec}}(E_{\theta_{enc}}(x), y) - x||_2^2$$

Добавляем дискриминатор

Пусть x и x' различаются только только атрибутом, например, фото человека в очках и без них, тогда хотим, чтобы их внутренние представления E(x) и E(x') были одинаковыми.



Auto-encoder feedback
Encoder adversarial feedback
Discriminator feedback

Predictions *Groundthruth*

Дискриминатор

$$P_{\theta_{dis}}(y \mid E(x))$$

$$\mathscr{L}_{dis}(\theta_{dis} \mid \theta_{enc}) = -\frac{1}{m} \sum_{(x,y) \in D} \log P_{\theta_{dis}}(y \mid E(x))$$

$$\mathscr{L}(\theta_{enc}, \theta_{dec} \mid \theta_{dis}) = \frac{1}{m} \sum_{(x,y) \in D} ||D_{\theta_{dec}}(E_{\theta_{enc}}(x), y) - x||_2^2 -$$

$$-\lambda_E \log P_{\theta_{dis}}(1 - y \mid E_{\theta_{enc}}(x))$$

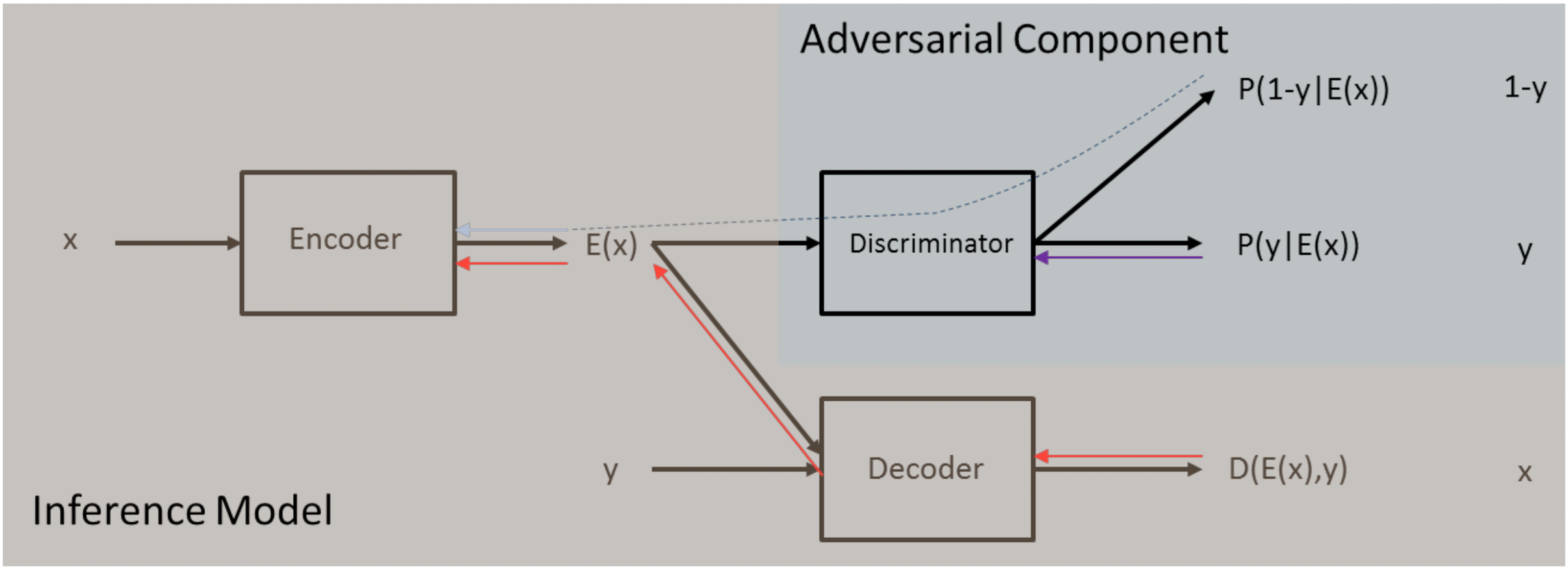
Обучение

$$\theta^*_{\text{enc}}, \theta^*_{\text{dec}} = \underset{\theta_{\text{enc}}, \theta_{\text{dec}}}{\operatorname{argmin}} \mathcal{L}(\theta_{\text{enc}}, \theta_{\text{dec}} | \theta^*_{\text{dis}}(\theta_{\text{enc}}))$$

Обновление весов:

$$\theta^{(t+1)}_{\text{dis}} = \theta^{(t)}_{\text{dis}} - \eta \nabla_{\theta_{\text{dis}}} \mathcal{L}_{\text{dis}}\left(\theta^{(t)}_{\text{dis}} \middle| \theta^{(t)}_{\text{enc}}, x^{(t)}, y^{(t)}\right)$$

$$[\theta^{(t+1)}_{\text{enc}}, \theta^{(t+1)}_{\text{dec}}] = [\theta^{(t)}_{\text{enc}}, \theta^{(t)}_{\text{dec}}] - \eta \nabla_{\theta_{\text{enc}}, \theta_{\text{dec}}} \mathcal{L}\left(\theta^{(t)}_{\text{enc}}, \theta^{(t)}_{\text{dec}} \middle| \theta^{(t+1)}_{\text{dis}}, x^{(t)}, y^{(t)}\right)$$



Auto-encoder feedback
Encoder adversarial feedback
Discriminator feedback

Predictions *Groundthruth*

Реализация

Энкодер

$$C_{16} - C_{32} - C_{64} - C_{128} - C_{256} - C_{512} - C_{512}$$

Декодер

$$C_{512+2n} - C_{512+2n} - C_{256+2n} - C_{128+2n} - C_{64+2n} - C_{32+2n} - C_{16+2n}$$

Дискриминатор - C_{512} с двумя полносвязными слоями размеров 512 и n

Ck - Convolution-BatchNorm-ReLU с k фильтрами

n - число атрибутов

Обучение

- λ_E увеличивали на 0.0001 на каждой эпохе в течение 500 000 итераций

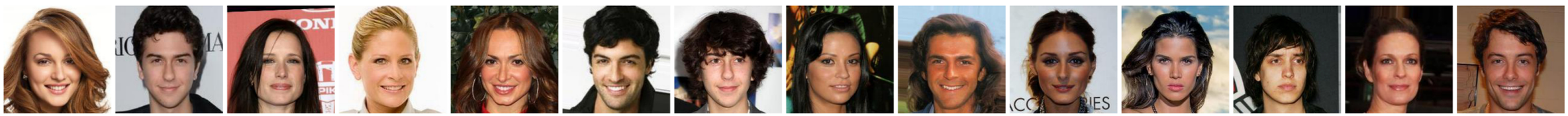
- Обучали на celebA

Результаты

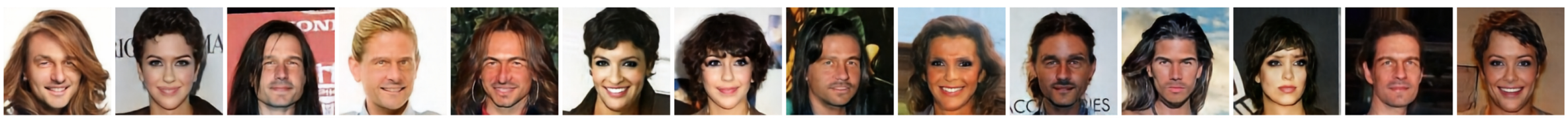
Model	Naturalness			Accuracy		
	Mouth	Smile	Glasses	Mouth	Smile	Glasses
Real Image	92.6	87.0	88.6	89.0	88.3	97.6
IcGAN AE	22.7	21.7	14.8	88.1	91.7	86.2
IcGAN Swap	11.4	22.9	9.6	10.1	9.9	47.5
FadNet AE	88.4	75.2	78.8	91.8	90.1	94.5
FadNet Swap	79.0	31.4	45.3	66.2	97.1	76.6

Результаты

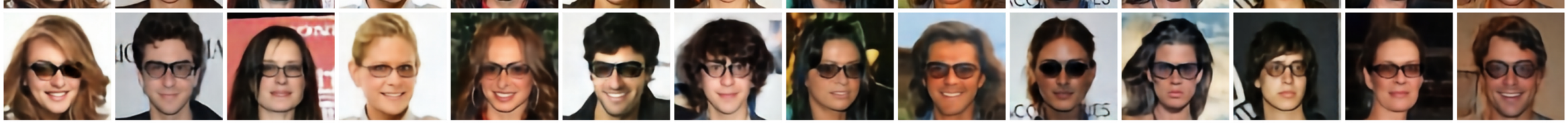
Original



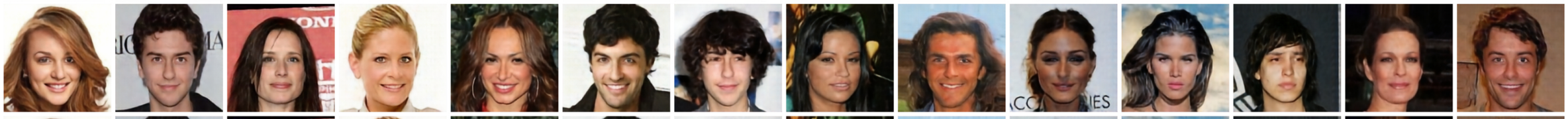
Gender



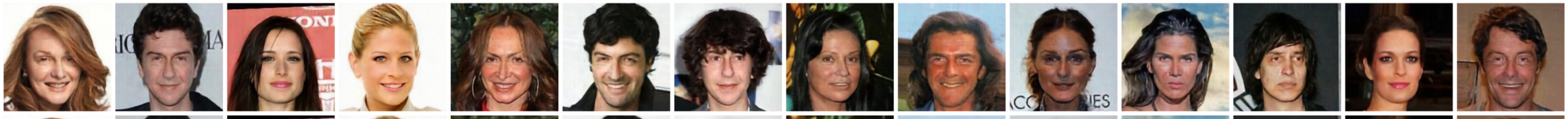
Glasses



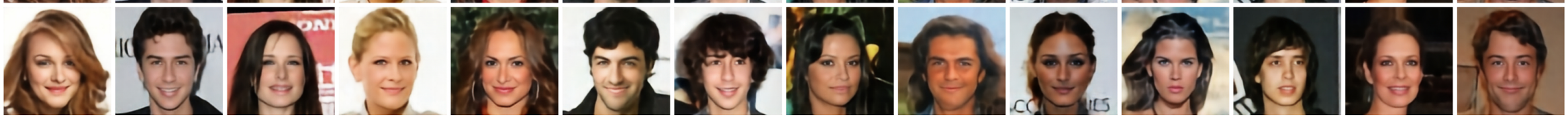
Eyes opening



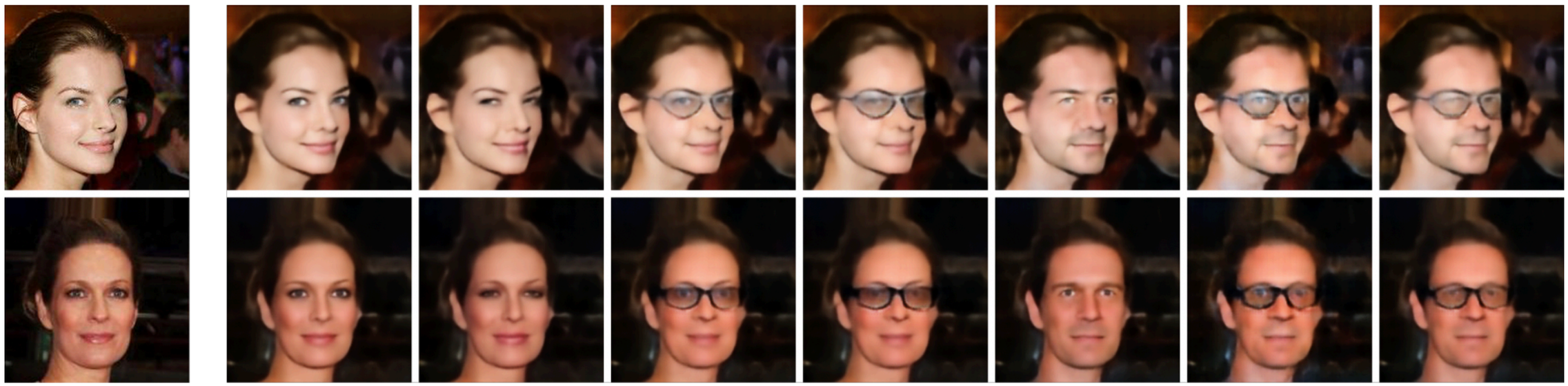
Age



Mouth opening



Результаты



Используемая литература

- A Beginner's Guide to Generative Adversarial Networks (GANs) https://skymind.ai/wiki/generative-adversarial-network-gan

- Conditional Generative Adversarial Nets https://arxiv.org/abs/1411.1784

- Invertible Conditional GANs for image editing https://arxiv.org/abs/1611.06355 https://github.com/Guim3/IcGAN

- Fader Networks: Manipulating Images by Sliding Attributes https://arxiv.org/abs/1706.00409 https://github.com/facebookresearch/FaderNetworks