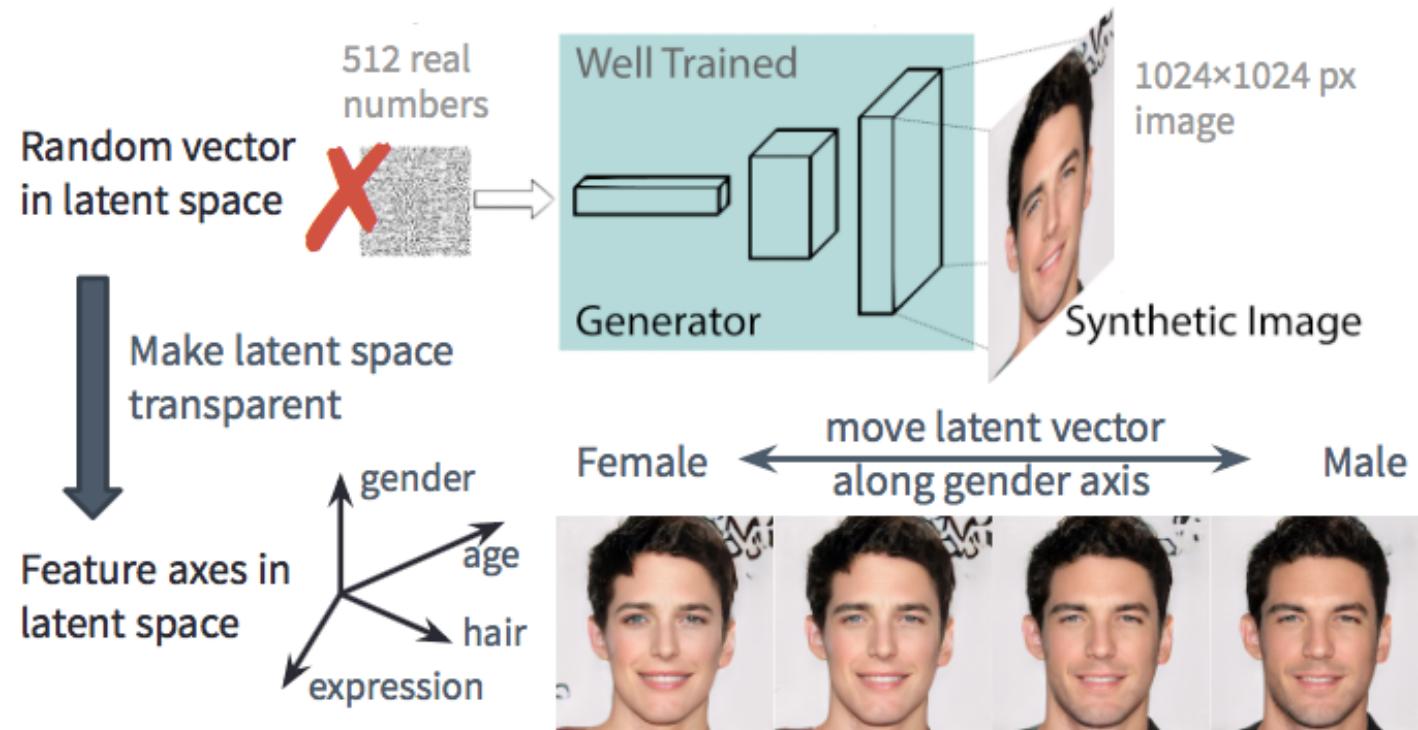


# Unsupervised Discovery of Interpretable Directions in the GAN Latent Space

Трус В.А. БПМИ172

# Мотивация для метода

Скрытое пространство (для GAN) – это векторное пространство, в котором находятся векторы, которые подаются на вход в генератор.



# Компоненты метода

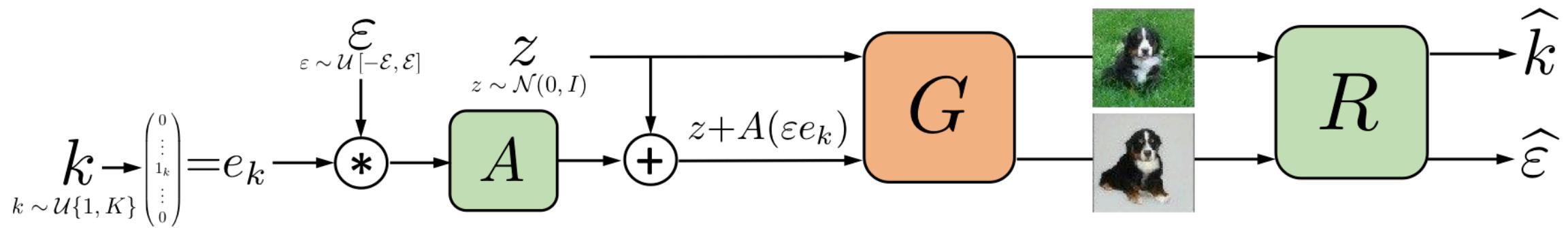
- $G: z \rightarrow I, z \in R^d$  - предобученный генератор
- $A \in R^{d \times K}$ ,  $d$  – размерность скрытого пространства,  $K$  – количество направлений
- $R$  получает на вход  $(G(z), G(z + A(\varepsilon e_k)))$ ,  $z \sim N(0, 1)$ ,
- $R(I_1, I_2) = (\hat{k}, \hat{\varepsilon})$ ,  $\hat{k}$  – это предсказание индекса направления,  $k \in \{1, \dots, K\}$ ,  $\hat{\varepsilon}$  – это предсказание сдвига  $\varepsilon$ ,  $e_k = (0, \dots, 1_k, \dots 0)$   
– выровненный по оси вектор
- $R: (I_1, I_2) \rightarrow (\{1, \dots, K\}, \mathbb{R})$

# ФУНКЦИЯ ПОТЕРЬ

$$\min_{A,R} \mathbb{E}_{z,k,\varepsilon} L(A, R) = \min_{A,R} \mathbb{E}_{z,k,\varepsilon} \left[ L_{cl}(k, \hat{k}) + \lambda L_r(\varepsilon, \hat{\varepsilon}) \right]$$

$L_{cl}(\cdot, \cdot)$  – кросс – энтропия,  $L_r(\cdot, \cdot)$  – MAE,  $\lambda = 0.25$

# Процесс обучения



# Детали обучения

Для реконструктора использовалась LeNet для MNIST и AnimeFaces и ResNet-18 для Imagenet и CelebA-HQ.

Количество каналов было установлено равным 6 (2 для MNIST).

*Распределения:*  $z \sim N(0, 1)$ ,  $k \sim U\{1, K\}$ ,  $\varepsilon \sim U[-6, 6]$

$\varepsilon$  – на практике использовалось  $\text{sign}(\varepsilon) \max(|\varepsilon|, 0.5)$

# Детали обучения

Spectral Norm GAN с Anime Faces: K = 128, с MNIST: K = 64

BigGAN: K = 120, ProgGAN: K = 200

Во всех экспериментах использовали Adam, для A и для R.

Learning rate = 0.0001. Было выполнено  $2 * 10^5$  градиентных шагов для Prog-GAN,  $10^5$  шагов для остальных моделей. Batch size = 128 для Spectral Norm GAN, 32 для BigGAN и 10 для ProgGAN. Все эксперименты были выполнены на NVIDIA Tesla v100.

# Выбор гиперпараметра A

- A – единичный линейный оператор, минус – часто высокие l2-нормы у столбцов
- A – линейный оператор с матрицей со столбцами единичной длины
- A – линейный оператор с ортонормированными столбцами матрицы

# Используемые метрики

- Reconstructor Classification Accuracy (RCA) – позволяет сравнить полученные методом направления со случайными/координатными осями. Необходимо устанавливать А случайной или единичной и не оптимизировать ее во время обучения.

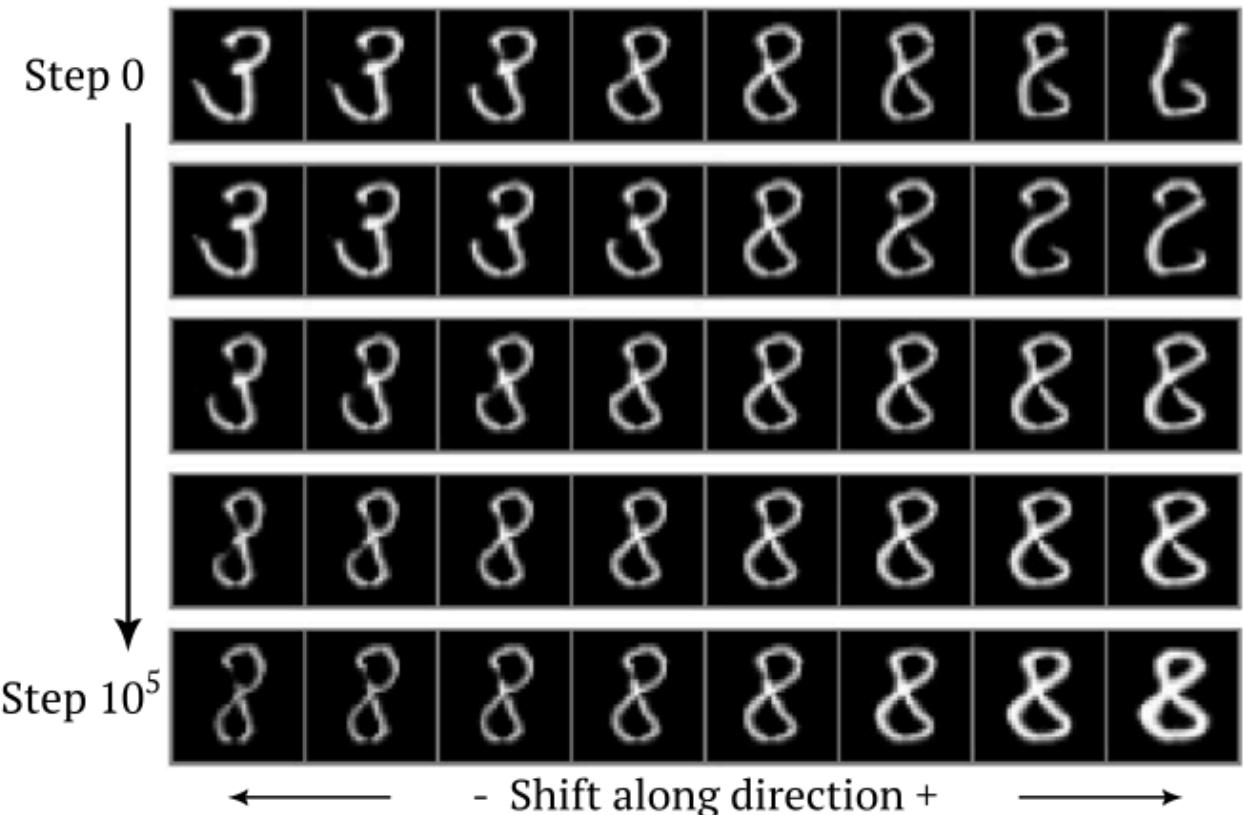
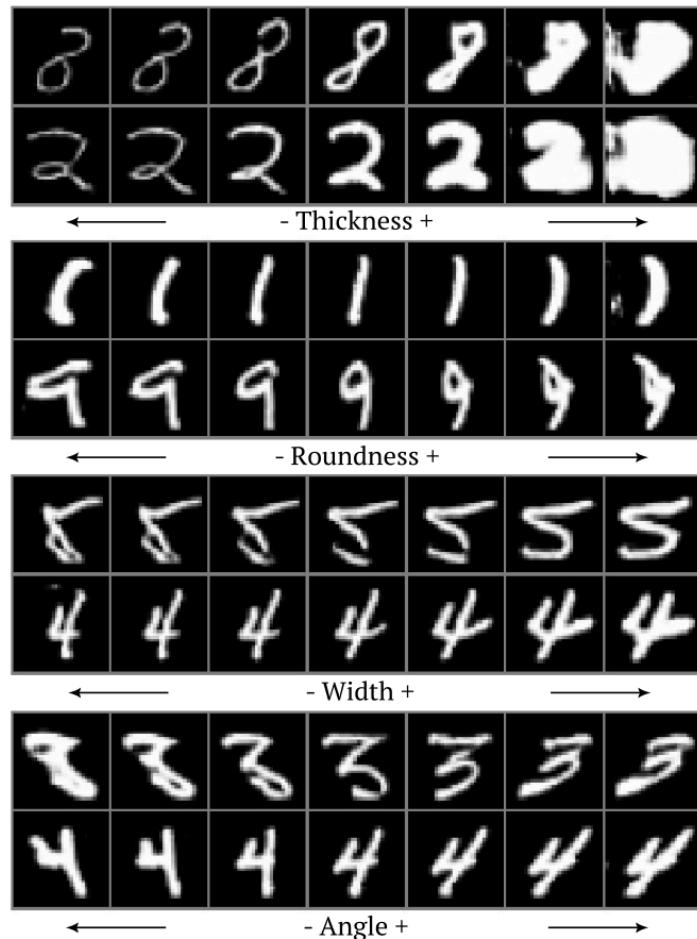
# Метрики

- *Individual interpretability (mean-opinion-score, MOS)* – человеческая оценка
- Для каждого эксперта семплируется 10 рандомных  $z$
- Для каждого эксперта рисуется график  $G(z+s^*h)$ , где варьировали  $s$  от -8 до 8, для всех  $z$  из прошлого шага

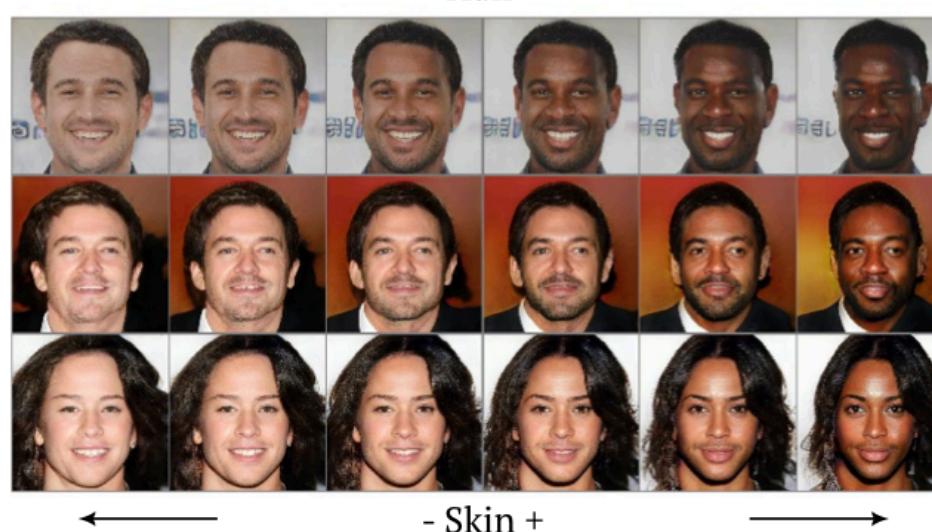
Задают вопросы:

1. Действует ли  $h$  последовательно?
2. Влияет ли оператор  $h$  на единственный фактор вариации, который легко интерпретировать?

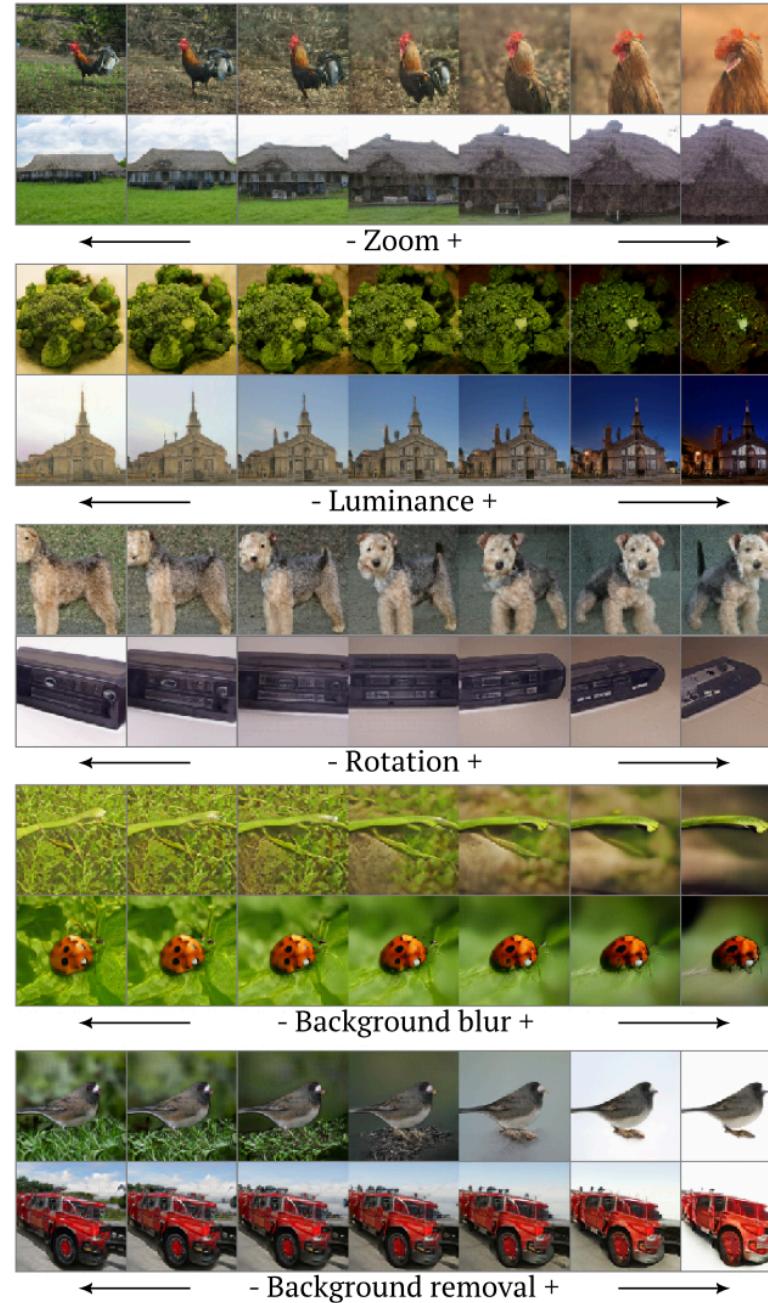
# Метрики



# Результаты



# Результаты



# Weakly-supervised saliency detection

Используется BigGan, генератору явно передается класс Imagenet:  $G(z, c)$ ,  $1 \leq c \leq 1000$ . За изменение прозрачности фона отвечает  $h_{bg}$ . Можно ввести маску с помощью соответствующего сдвинутого изображения.

$$\text{Mask}(G(z, c)) = [ G(z + h_{bg}, c) < \theta ]$$



*Figure 10.* Segmentation masks for BigGAN samples used to train a saliency detection model. *Line 1:* original samples  $G(z)$ ; *Line 2:* samples with reduced background  $G(z + h_{bg})$ ; *Line 3:* generated binary masks obtained by thresholding.

# Количество скрытых направлений

*Table 3.* Number of directions  $K$  ablation for Spectral Norm GAN pretrained on MNIST dataset.

metrics	$K = 16$	32	64	128
MOS	0.5	0.58	0.47	0.46
MOS (absolute)	8	19	30	59
RCA	0.98	0.95	0.88	0.79

*Table 4.* Number of directions  $K$  ablation for BigGAN.

metrics	$K = 15$	30	60	90	120
MOS	0.3	0.3	0.38	0.75	0.69
MOS (absolute)	5	9	23	68	83
RCA	0.99	0.98	0.92	0.9	0.85

*Table 5.* Number of directions  $K$  ablation for Spectral Norm GAN pretrained on MNIST dataset.

metrics	$\lambda = 0$	0.125	0.25	0.5	2
MOS	0.27	0.35	0.47	0.42	0.25
RCA	0.88	0.90	0.88	0.87	0.75

# Выводы

- 1. Авторы статьи предложили первый unsupervised подход по обнаружению семантически значимых направлений в скрытом пространстве GAN.
- 2. Для нескольких известных генераторов обнаружили нетривиальные и практически важные направления.
- 3. Привели пример практической пользы для удаления фона для weakly-supervised saliency detection.

# Вопросы

- Объясните понятия смещателя и реконструктора, объясните процесс обучения метода.
- Какие метрики использовали авторы статьи?
- Как подбирались параметры К и А?