

# Adversarially Learned One-Class Classifier for Novelty Detection

Арсений Белков B01-901

# Novelty Detection

- Novelty Detection - задача обнаружения данных (novelty class), которые отличаются от тех, что используются для обучения (target class).
- Эту задачу можно решать как one class классификацию

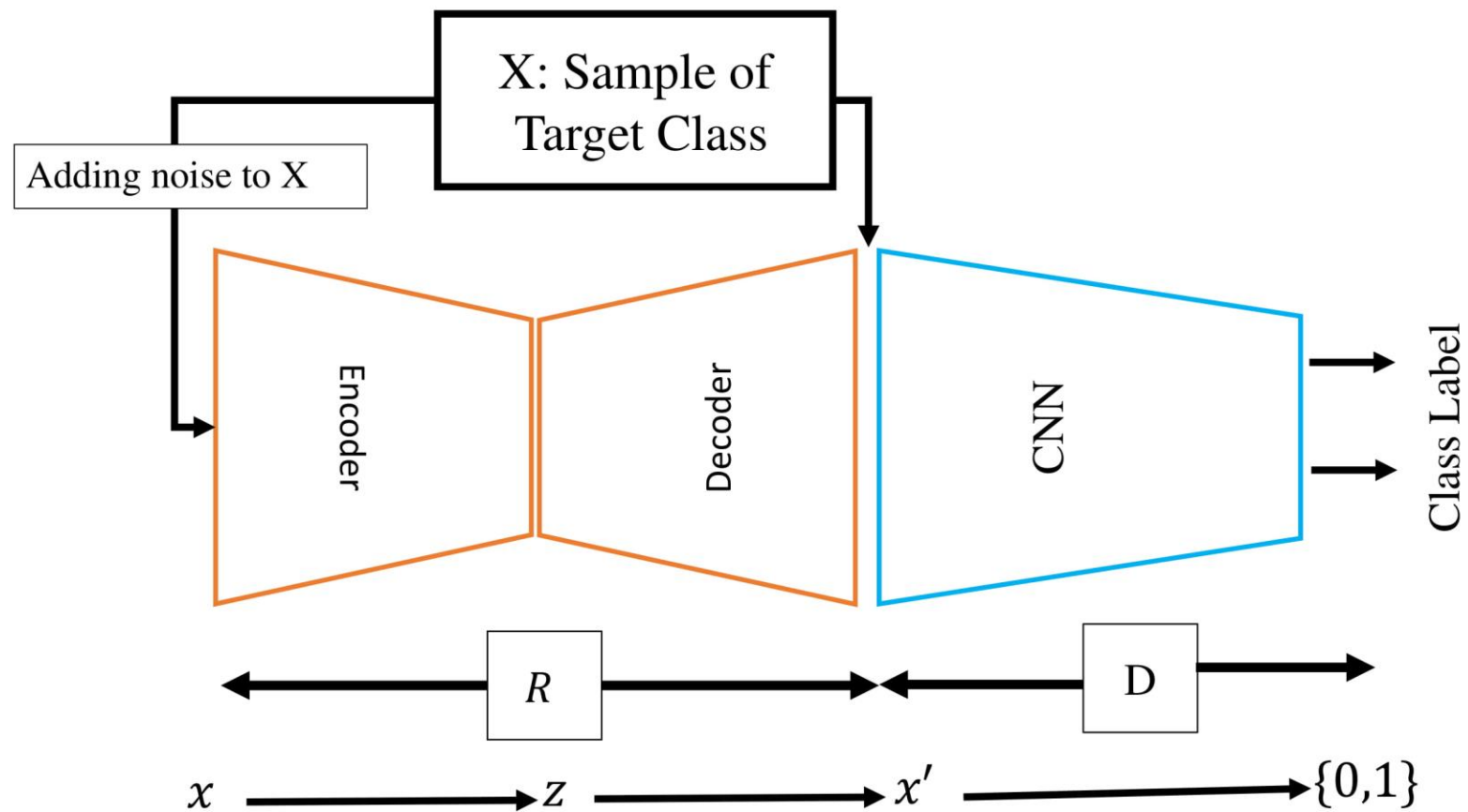
Target class



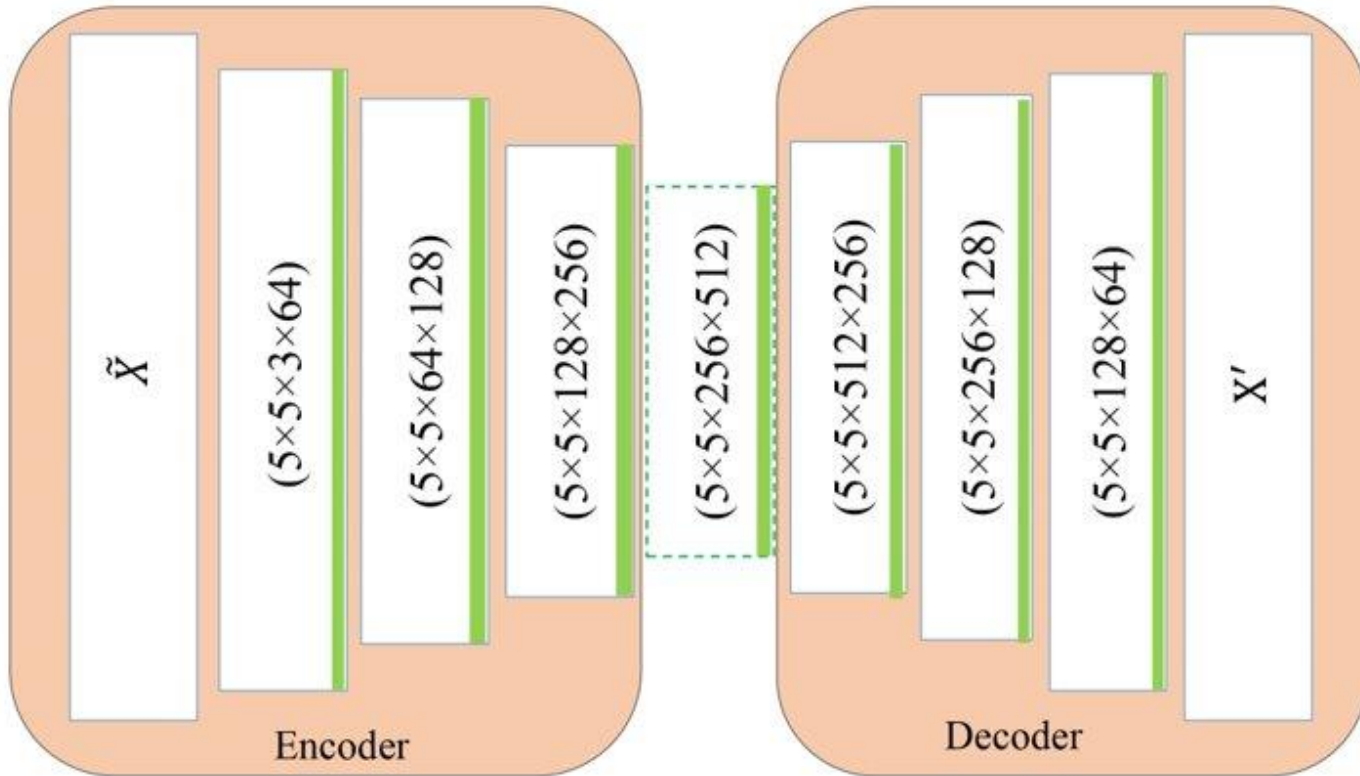
Novelty class



# Предложенный метод



# R Network



## Encoder:

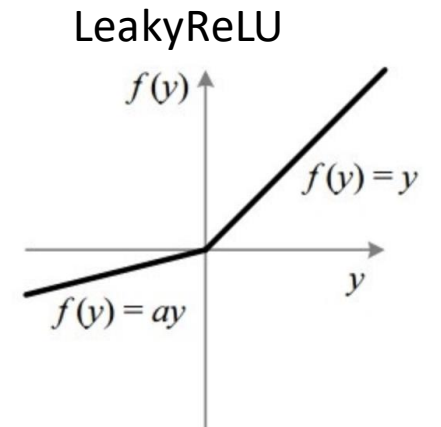
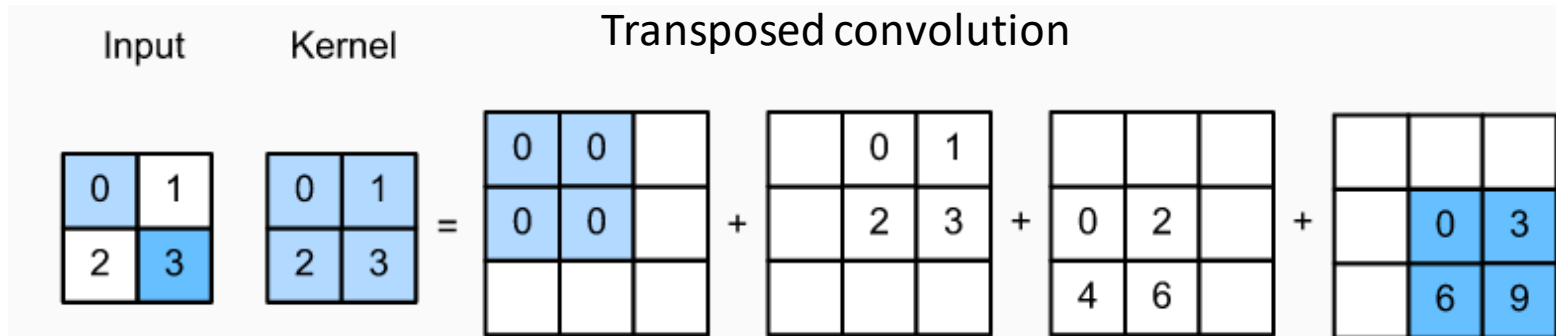
Сверточная сеть с BatchNorm2d и LeakyReLU в качестве функции активации.

## Decoder:

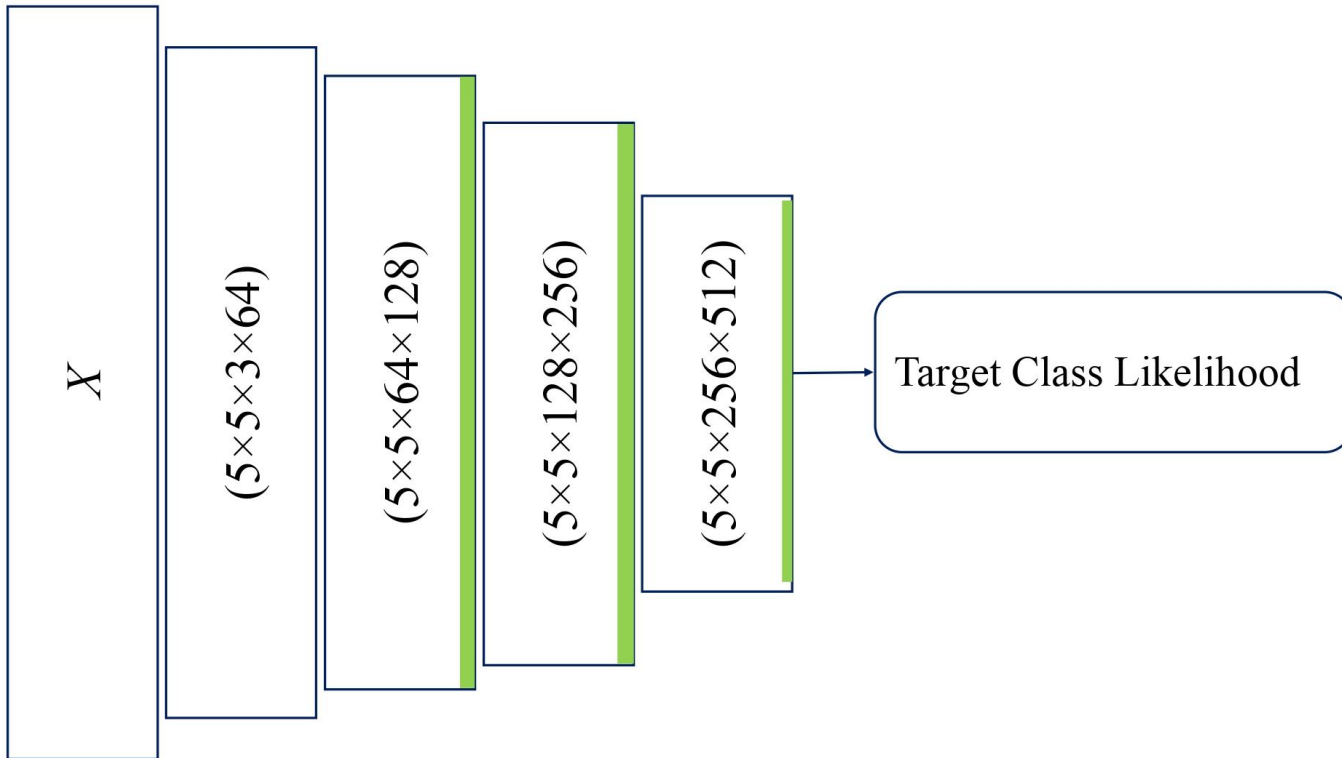
Сверточная сеть с transposed convolution, BatchNorm2d, LeakyReLU.

Refinement loss:

$$\mathcal{L}_{\mathcal{R}} = \|X - X'\|^2$$



# D Network



- CNN:
- Сверточная сеть с LeakyReLU и BatchNorm2d.
- Классификатор:
- Полносвязная сеть с Sigmoid

# Обучение

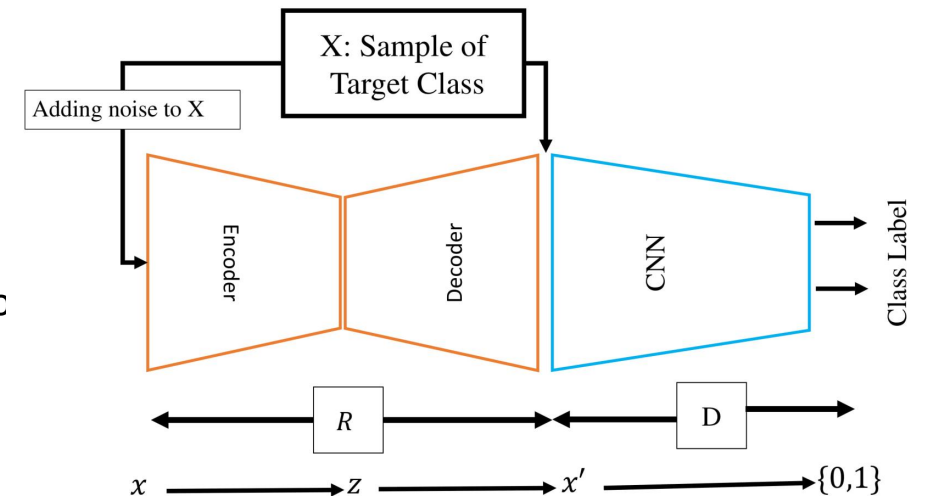
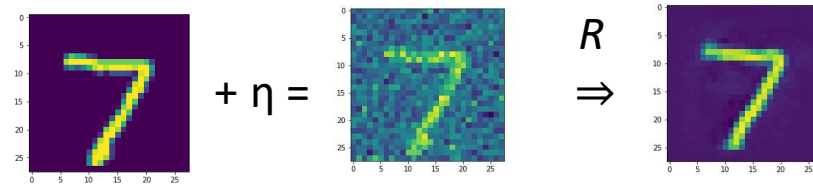
- Из данных сэмплируется  $X \sim p_t$
- Создается копия и зашумляется  $\tilde{X} = (X \sim p_t) + (n \sim \mathcal{N}(0, \sigma^2 \mathbf{I}))$
- Восстанавливаем  $X$ :  $X' = R(\tilde{X})$
- Обучаются  $D$  и  $R$ :  $\min_R \max_D \left( \mathbb{E}_{X \sim p_t} [\log(D(X))] + \mathbb{E}_{\tilde{X} \sim p_t + \mathcal{N}_\sigma} [\log(1 - D(R(\tilde{X})))] + \lambda \|X - R(\tilde{X})\|^2 \right)$
- 1) Обучаем  $D$  на  $X$  и  $\tilde{X}$ , помеченных как **1** и **0** соответственно:  $\min_D \left[ -\mathbb{E}_{X \sim p_t} \log D(X) - \mathbb{E}_{\tilde{X} \sim p_t + \mathcal{N}_\sigma} \log(1 - D(R(\tilde{X}))) \right]$
- 2) Обучаем  $R$  на  $\tilde{X}$ :  $\min_R \mathbb{E}_{\tilde{X} \sim p_t + \mathcal{N}_\sigma} [\log(1 - D(R(\tilde{X})))] \implies \min_R \left[ -\mathbb{E}_{\tilde{X} \sim p_t + \mathcal{N}_\sigma} \log D(R(\tilde{X})) \right]$

Reconstruction loss:  $\mathcal{L}_R = \|X - X'\|^2$

$$\min_R \mathbb{E}_{\tilde{X} \sim p_t + \mathcal{N}_\sigma} \left( -\log D(R(\tilde{X})) + \lambda \|X - R(\tilde{X})\|^2 \right)$$

- Тренировка останавливается, когда  $\|X - X'\|^2 < \rho$ ,  $\rho$  - малое









позитивное число



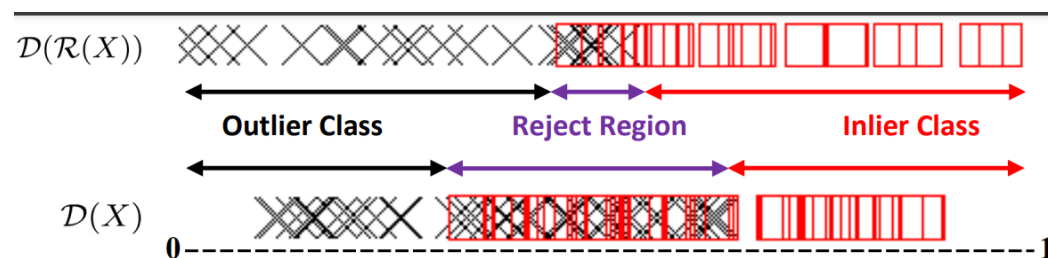
# Как это работает?

- После совместной тренировки сетей  $R$  и  $D$ :
  - $R$  была натренирована как denoising autoencoder, следовательно она восстанавливает  $\tilde{X} = (X \sim p_t) + (n \sim \mathcal{N}(0, \sigma^2 \mathbf{I}))$  в  $X' \sim p_t$
  - Так как novelty class  $\hat{X}$  присутствует в данных в очень малом количестве или не присутствует вовсе, то  $R$  не сможет восстановить (испортит) данные из этого класса, т.к. не была на них обучена.
  - Так как  $D$  натренирован принимать данные из  $p_t$  как true label, то можно ожидать, что испорченные  $\hat{X}$  он будет детектировать как выбросы.

$$\text{OCC}_2(X) = \begin{cases} \text{Target Class} & \text{if } \mathcal{D}(\mathcal{R}(X)) > \tau \\ \text{Novelty (Outlier)} & \text{otherwise.} \end{cases}$$

	Noisy Inlier Samples		Outlier Samples	
$X$				
$\mathcal{R}(X)$				
$\mathcal{D}(X)$	0.75	0.72	0.53	0.27
$\mathcal{D}(\mathcal{R}(X))$	<b>0.85</b>	<b>0.91</b>	0.25	0.10

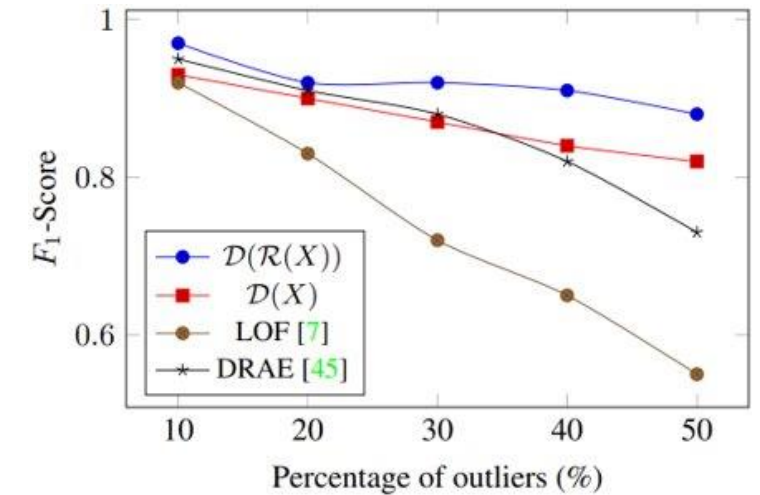
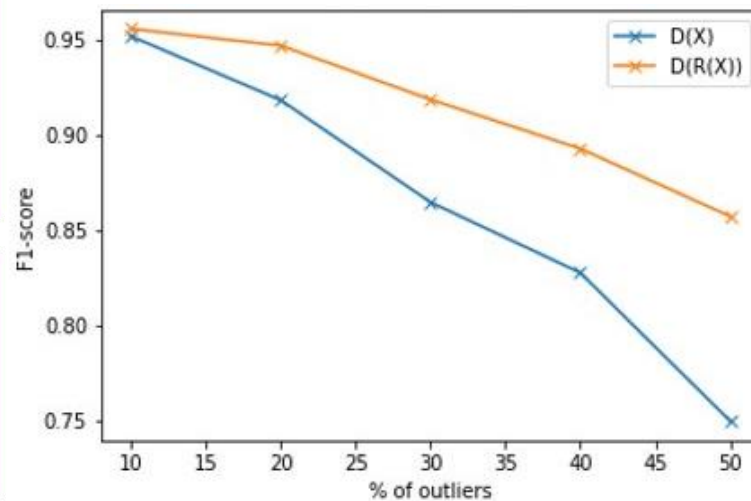
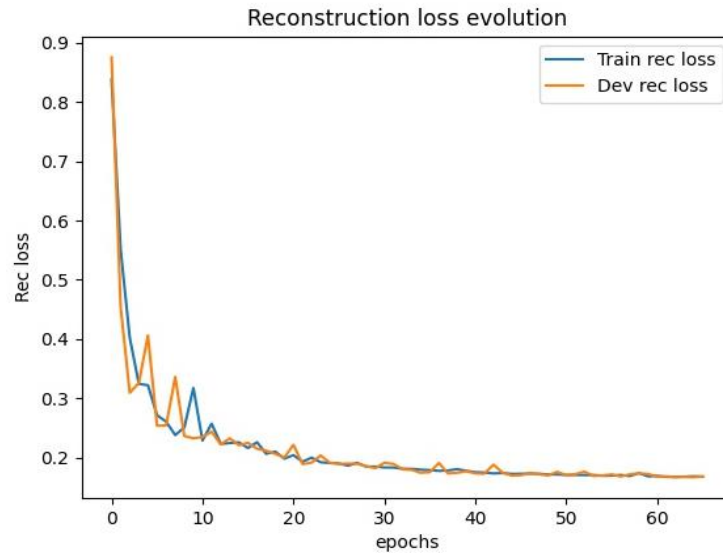
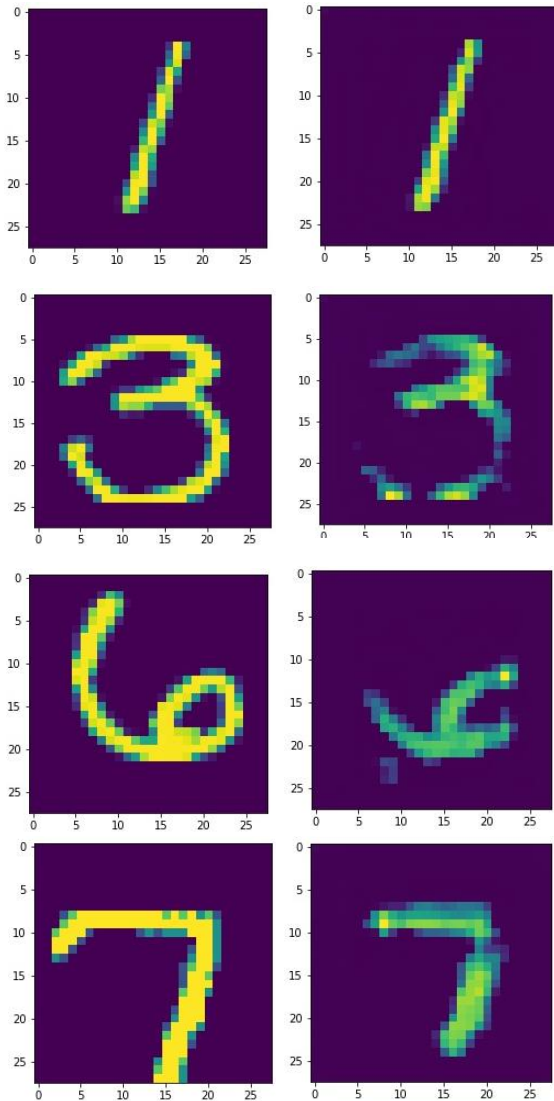
$$\mathcal{D}(\mathcal{R}(X \sim p_t)) - \mathcal{D}(\mathcal{R}(\hat{X} \sim p_{\gamma})) > \mathcal{D}(X \sim p_t) - \mathcal{D}(\hat{X} \sim p_{\gamma})$$





# Результаты

Подробнее: [https://github.com/Ars235/Novelty\\_Detection](https://github.com/Ars235/Novelty_Detection)

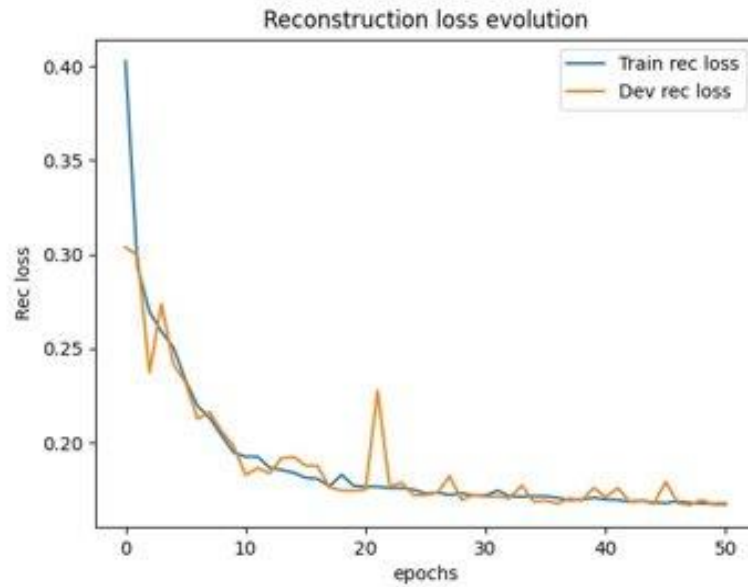


- Преимущества
  - Устойчива к изменению количества выбросов и target классов
  - Можно обучить в отсутствии novelty класса
  - Отсутствует mode collapse
- Недостатки
  - Тяжело обучается

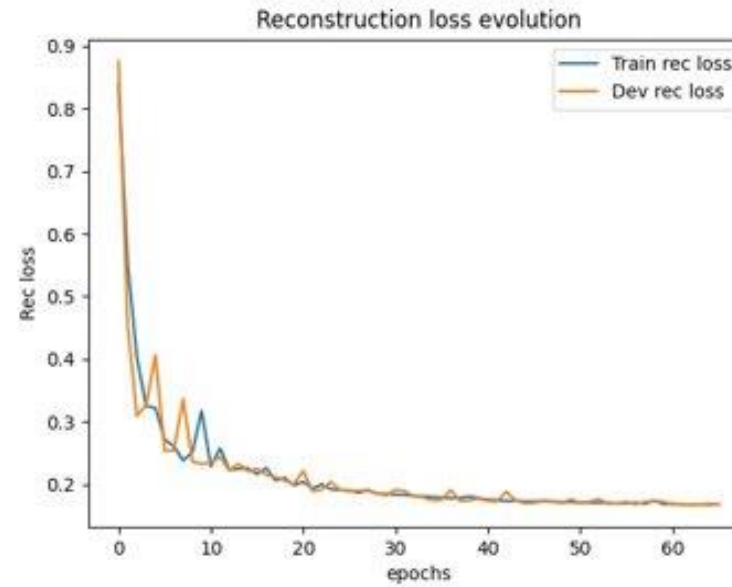


# Applying wasserstein loss

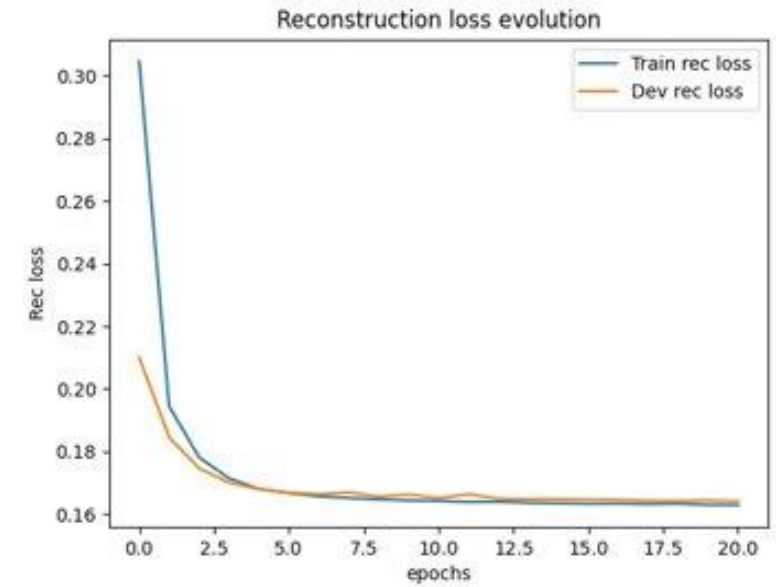
GAN loss,  $lr = 0.0001$



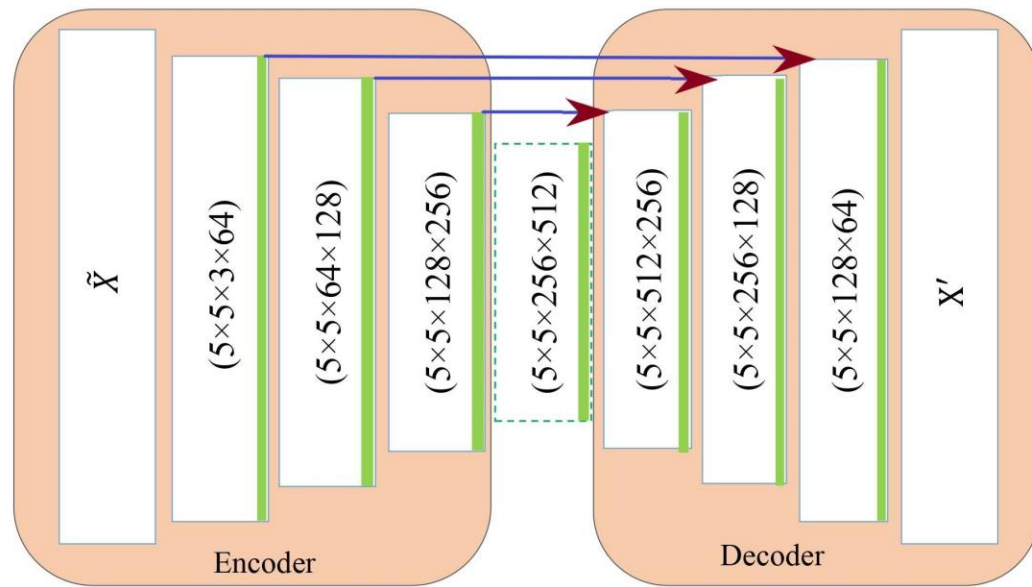
GAN loss,  $lr = 0.001$



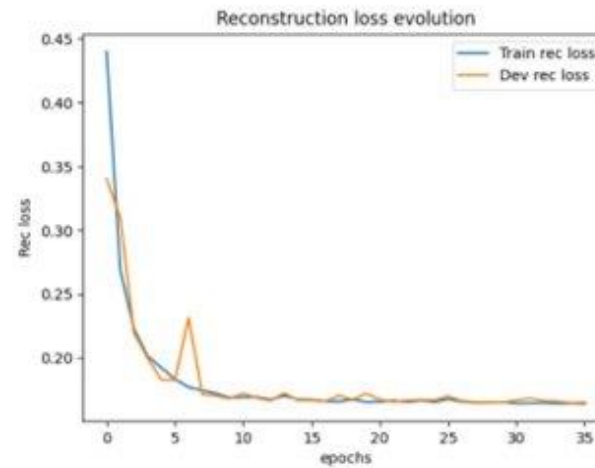
Wasserstein loss,  $lr = 0.0001$



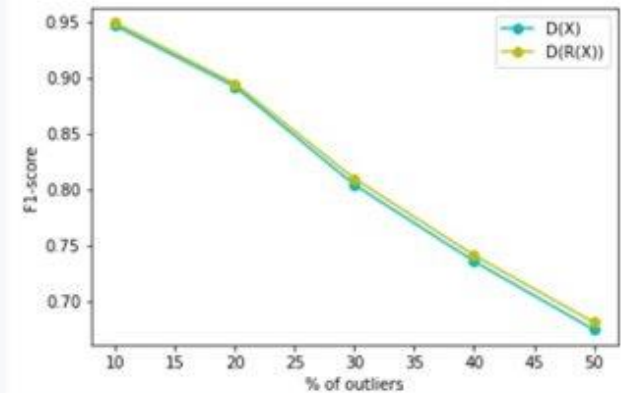
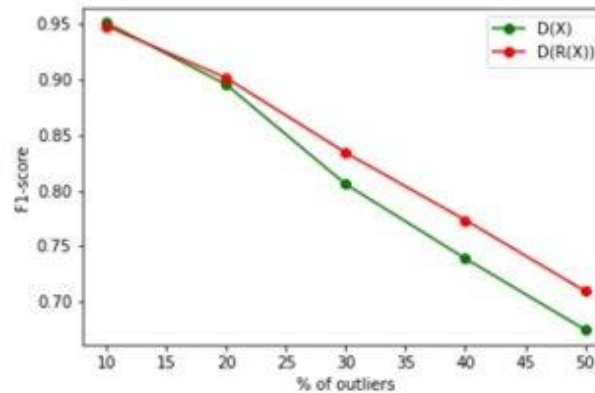
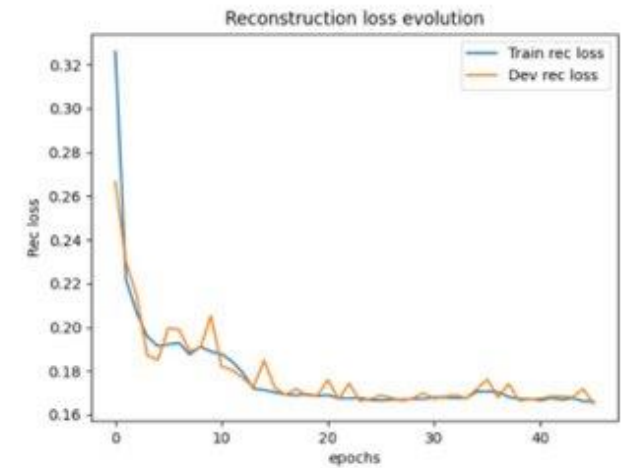
# Applying skip connections



Concatenation



Residual connections



# Литература

- M. Sabokrou, M. Khalooei, M. Fathy, E. Adeli: Adversarially Learned One-Class Classifier for Novelty Detection
- Martin Arjovsky, Soumith Chintala, and Léon Bottou: Wasserstein GAN
- Ian Goodfellow et al.: Generative Adversarial Networks