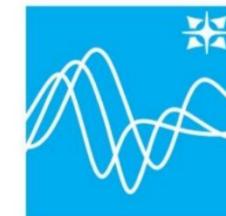


# Сверточные нейронные сети в задачах компьютерного зрения



Кафедра  
технологий  
проектирования  
сложных  
технических  
систем

# План лекции

- Обзор задач технического зрения
- Обзор технологии переноса обучения и самообучения
- Задача верификации

# Основные направления в CV

## Классификация

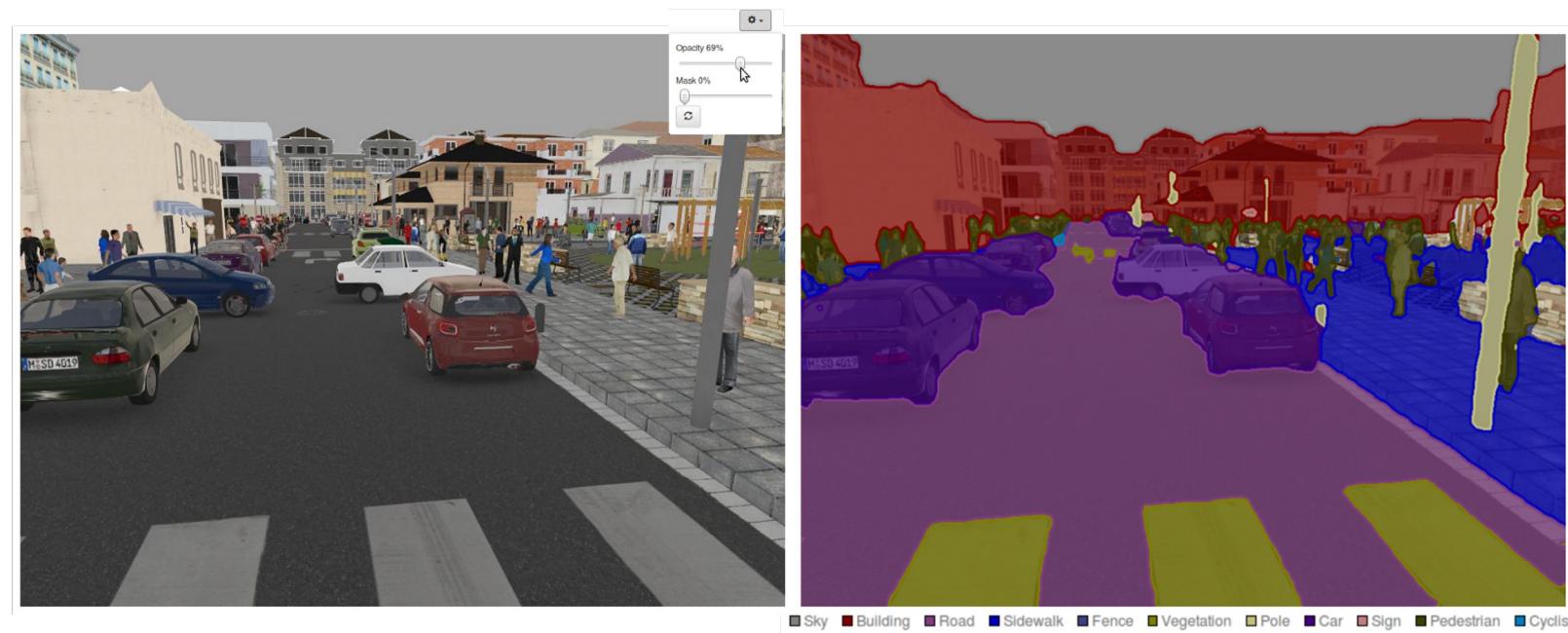
Решается задача  
принадлежности  
изображения одному из  
классов



# Основные направления в CV

## Сегментация

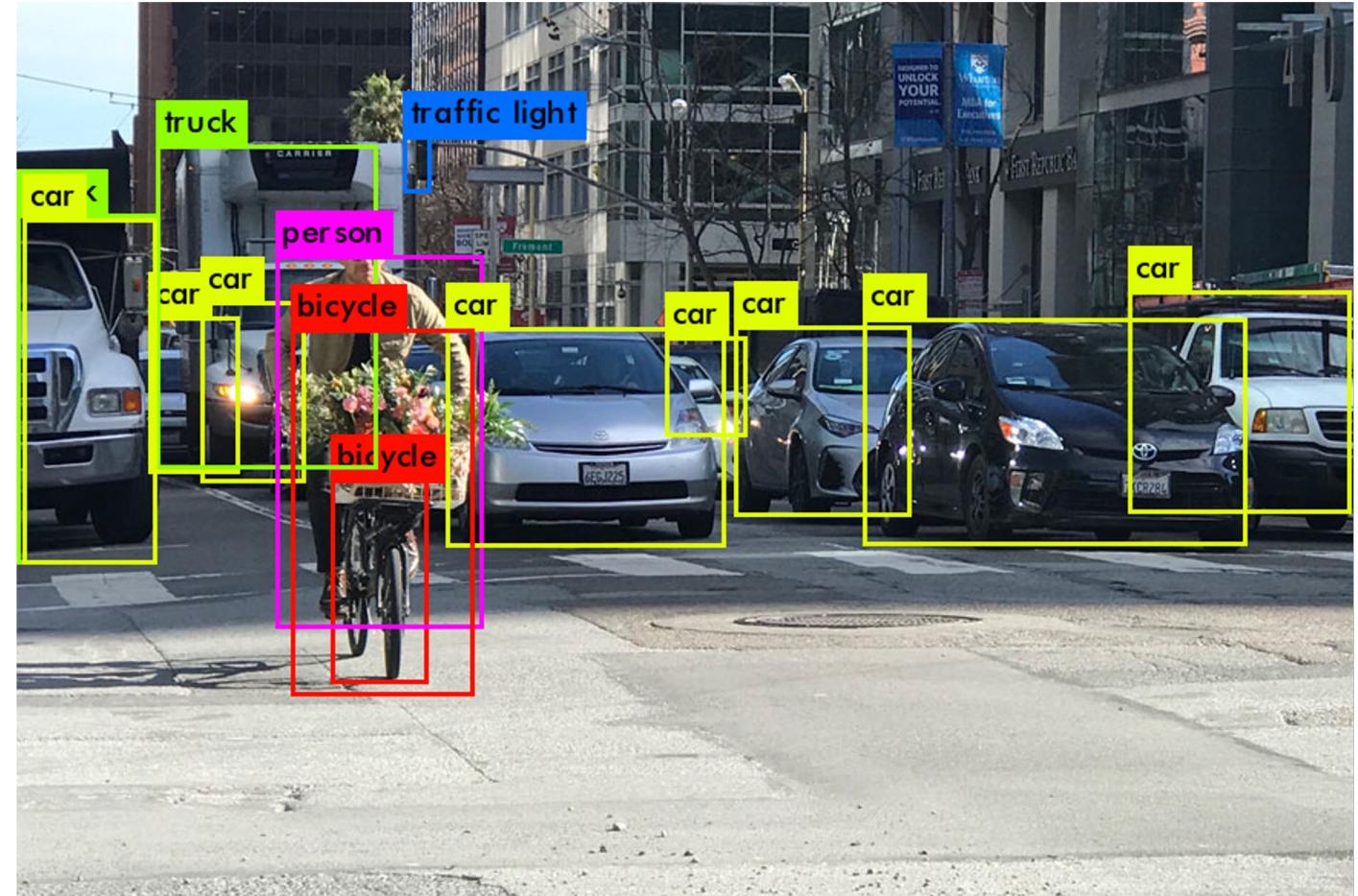
Решается задача  
классификации, но  
попиксельной



# Основные направления в CV

## Детекция

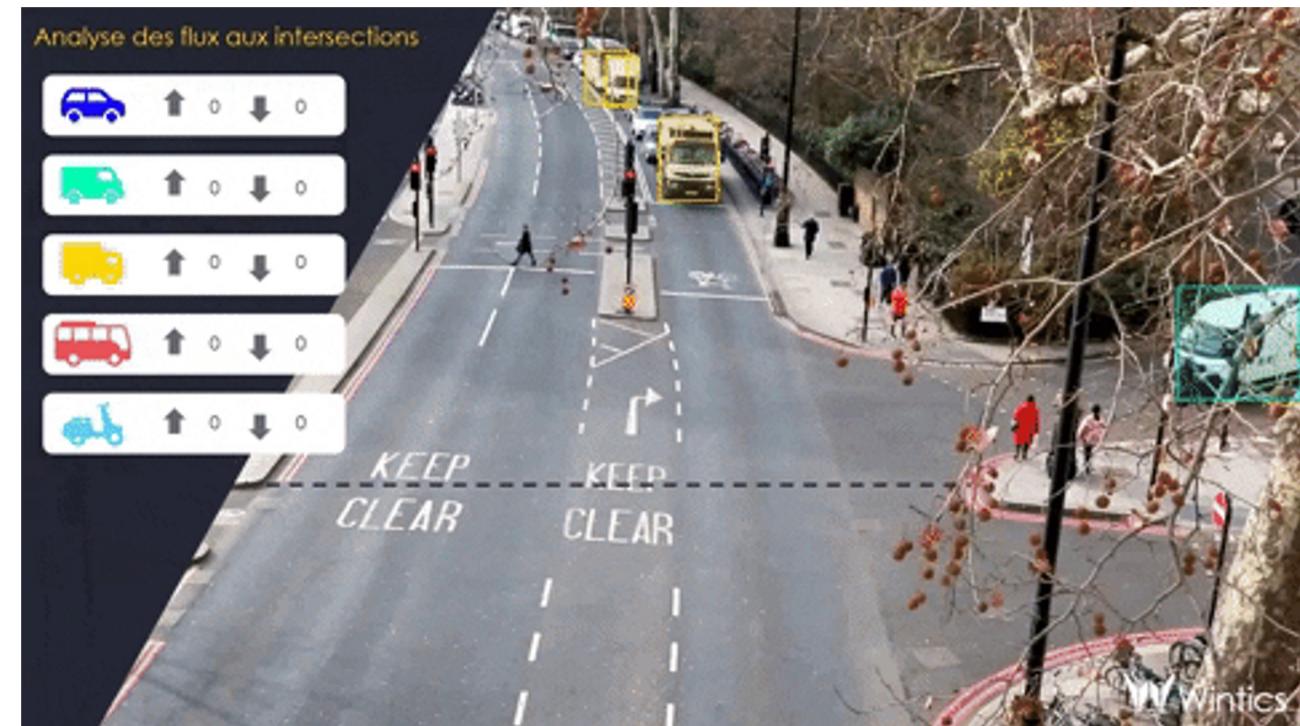
Решается задача  
распознавания - найти и  
классифицировать



# Основные направления в CV

Трекинг и видео-аналитика

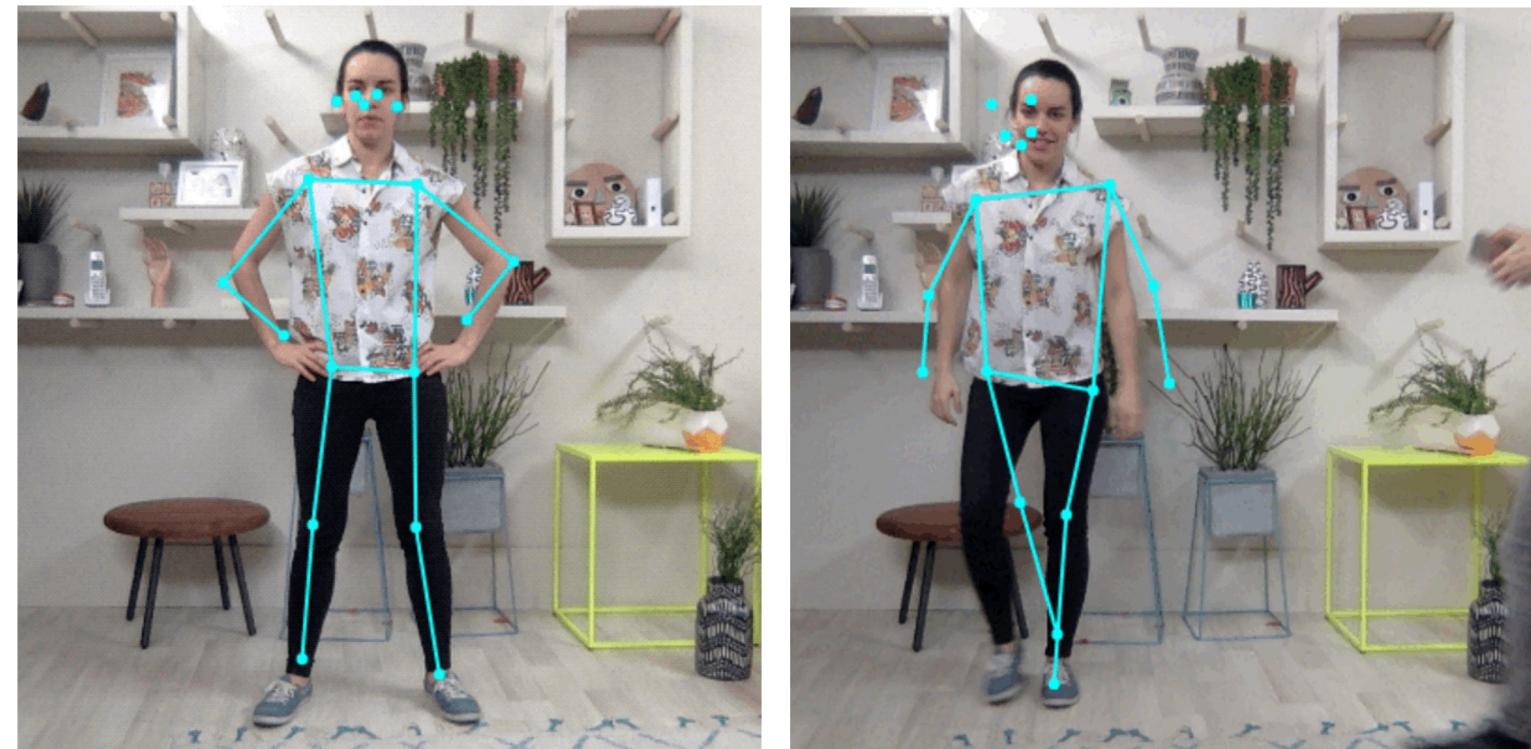
Распознавание объектов в видеопотоке



# Основные направления в CV

Оценка позы и взгляда человека

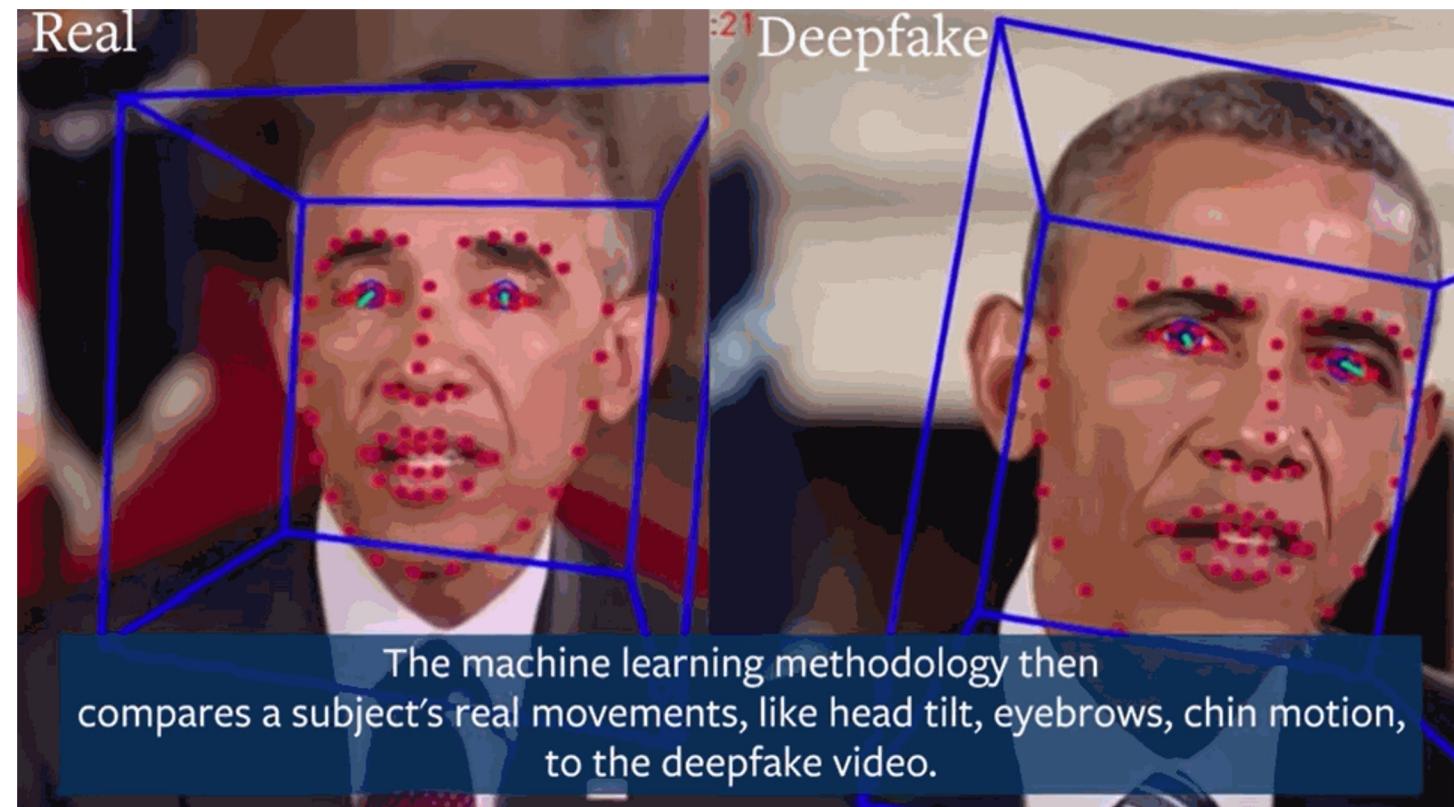
Решается задача  
распознавания ключевых  
точек, которые  
описывают позы  
человека, его положение,  
ориентацию в  
пространстве



# Основные направления в CV

## Биометрия

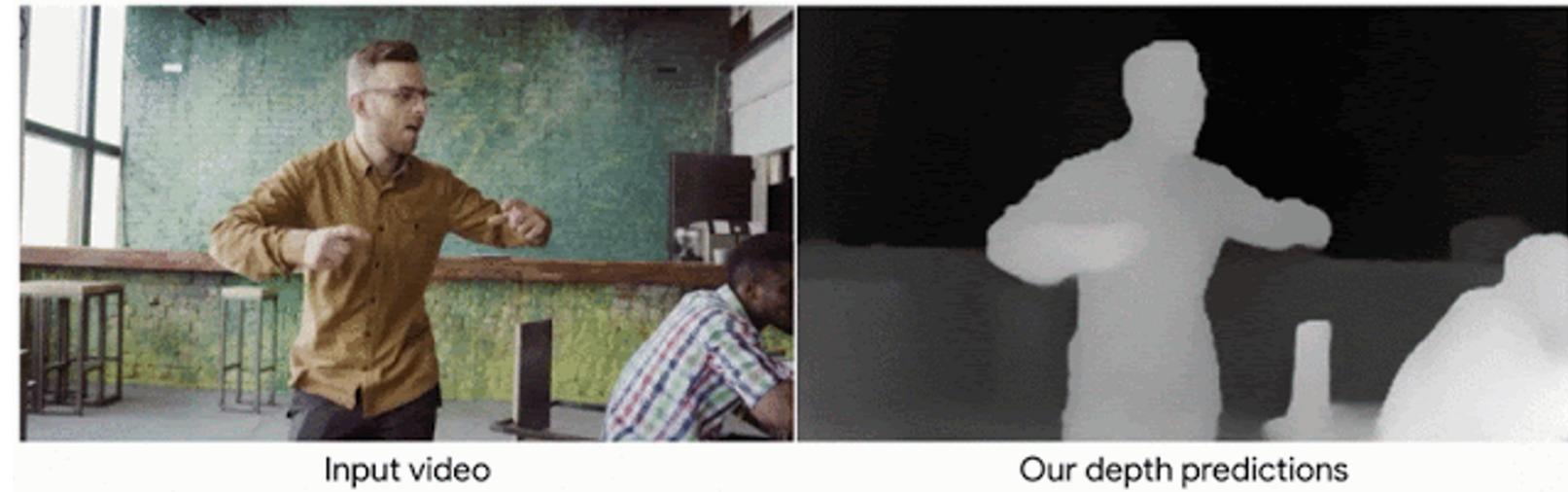
Решается задача классификации на изображении или видеопотоке: фейк или нет



# Основные направления в CV

Построение карты глубины

Решается задача  
построения карты  
расстояний до  
наблюдаемых объектов



# Основные направления в CV

Улучшение и восстановление изображений

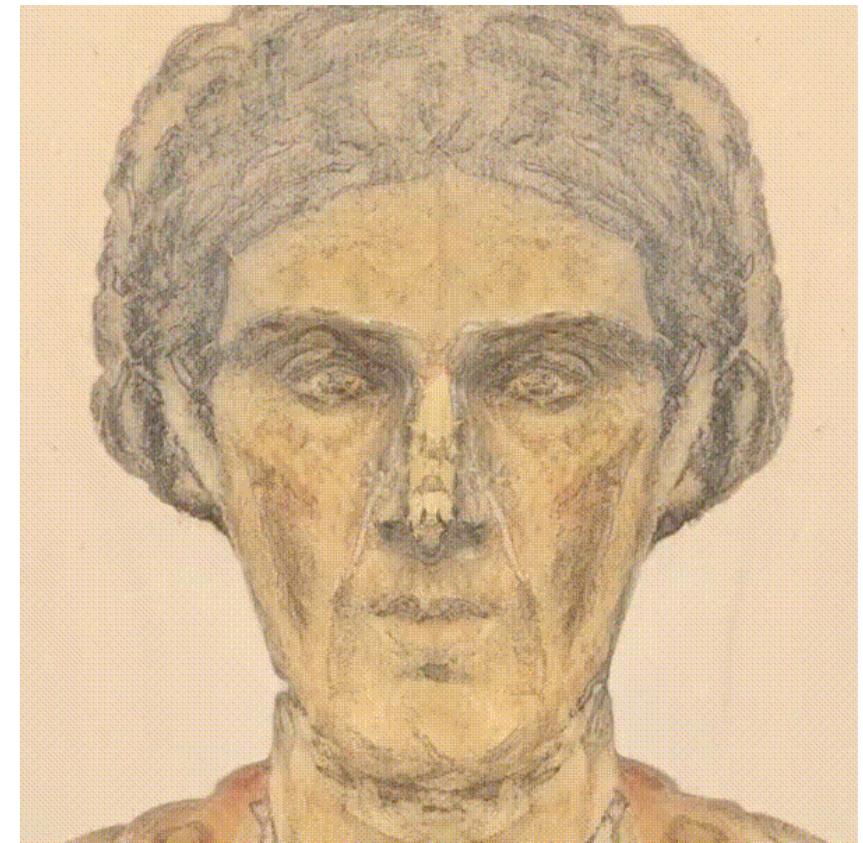
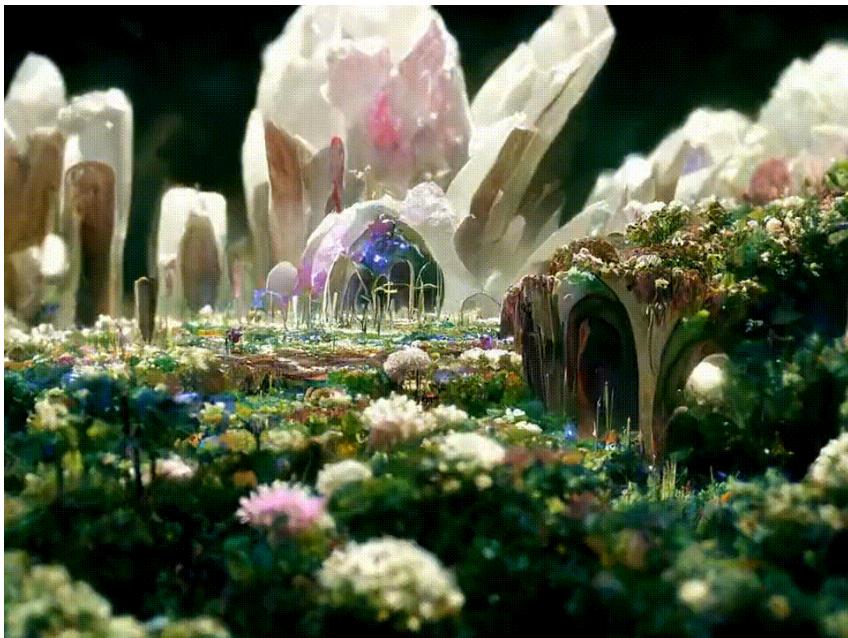
Поиск дефектов и  
коррекция изображения,  
преобразование стилей



# Основные направления в CV

Генерация изображений

Синтез новых  
изображений



# Основные направления в CV

Рендеринг

VR и нейро рендеринг



# Основные направления в CV

## Построение 3D Объектов

Решается задача  
построения объемного  
объекта (obj-файл) по  
двумерному  
изображению



# Типы задач обучения алгоритмов

— обучение по размеченным данным (Supervised Learning / SL) —

обучающая выборка состоит из пар  $(x, y)$ , где  $x$  – описание объекта,  $y$  – его метка, и необходимо обучить модель  $y=f(x)$ , которая по описаниям получает метки.

— обучение с частично размеченными данными (Semi-

Supervised Learning / SSL) — обучающая выборка состоит из данных с метками и без меток (последних, как правило, существенно больше), необходимо также обучить модель  $y=f(x)$ , но здесь может помочь информация о том, как объекты располагаются в пространстве описаний, см. рис. 1.

— обучение по неразмеченным данным (Unsupervised Learning / UL) —

даны только объекты (без меток), необходимо эффективно описать, как они располагаются в пространстве описаний.\*

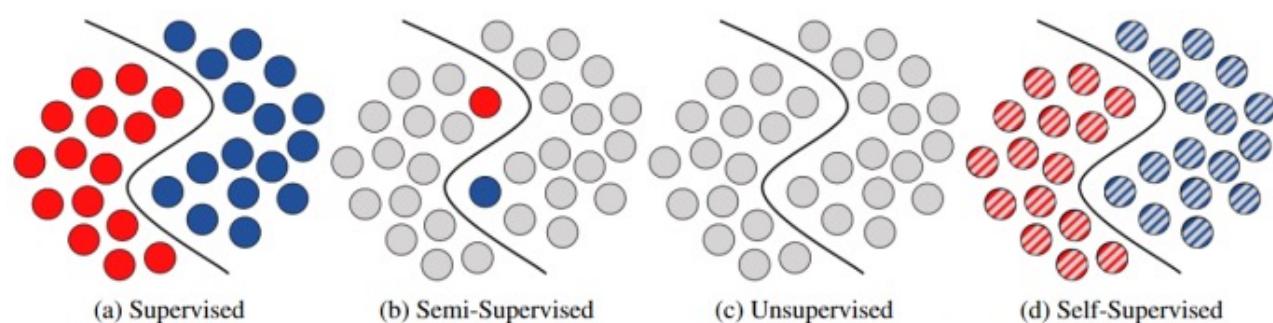
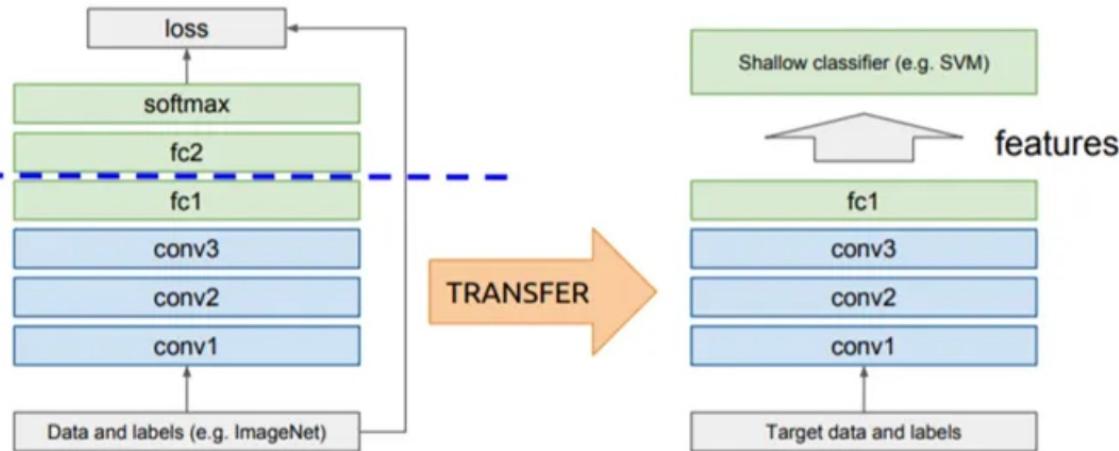


Figure 2: Illustrations of the four presented deep learning strategies - The red and dark blue circles represent labeled data points of different classes. The light grey circles represent unlabeled data points. The black lines define the underlying decision boundary between the classes. The striped circles represent datapoints which ignore and use the label information at different stages of the training process.

<https://arxiv.org/pdf/2002.08721.pdf>

# Перенос обучения

transfer learning



**Предварительная задача (pretext task)** – задача с искусственно созданными метками (псевдо-метками),

**Псевдо-метки (pseudo labels)** – метки, которые получаются автоматически, без ручной разметки,

**Последующая задача (downstream task)** – задача на которой проверяют качество полученных представлений.

**Самообучение (self-supervised)** – это направление в глубоком обучении, которое стремится сделать глубокое обучение процедурой предварительной обработки данных\*

# Подходы в самообучении

## Предсказание контекста

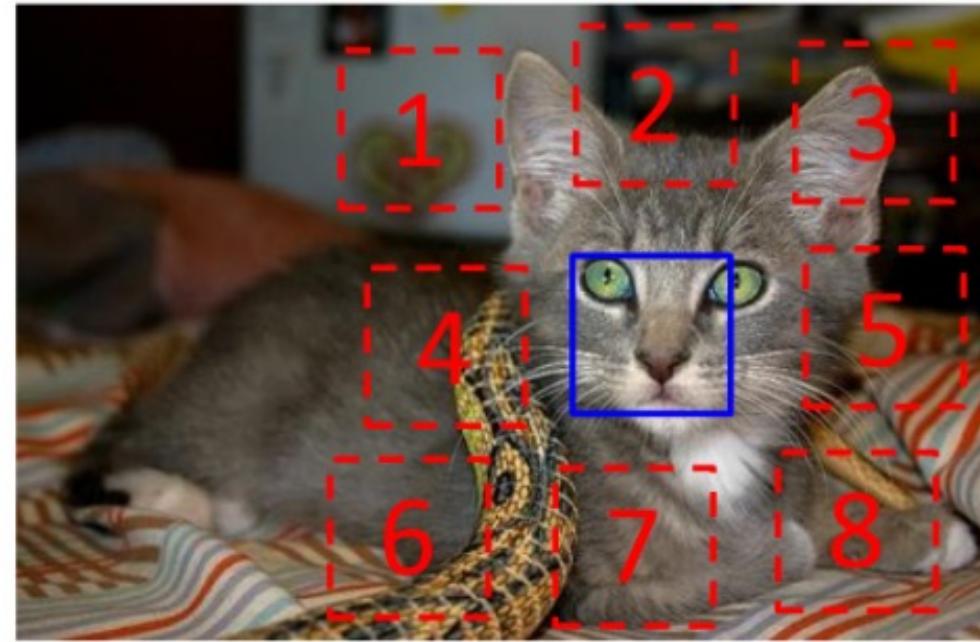


Figure 2. The algorithm receives two patches in one of these eight possible spatial arrangements, without any context, and must then classify which configuration was sampled.

# Подходы в самообучении

## Определение поворота



Figure 1: Images rotated by random multiples of 90 degrees (e.g., 0, 90, 180, or 270 degrees). The core intuition of our self-supervised feature learning approach is that if someone is not aware of the concepts of the objects depicted in the images, he cannot recognize the rotation that was applied to them.

# Подходы в самообучении

## Образец

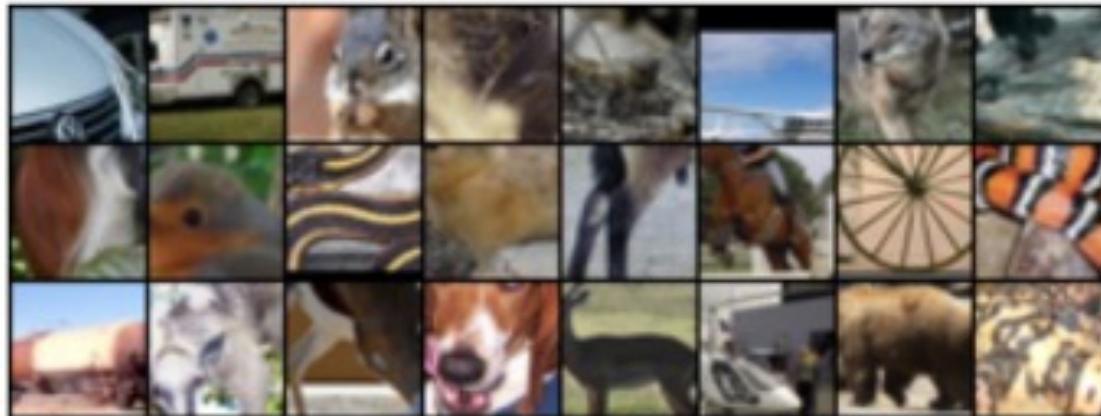


Fig. 1. Exemplary patches sampled from the STL unlabeled dataset which are later augmented by various transformations to obtain surrogate data for the CNN training.

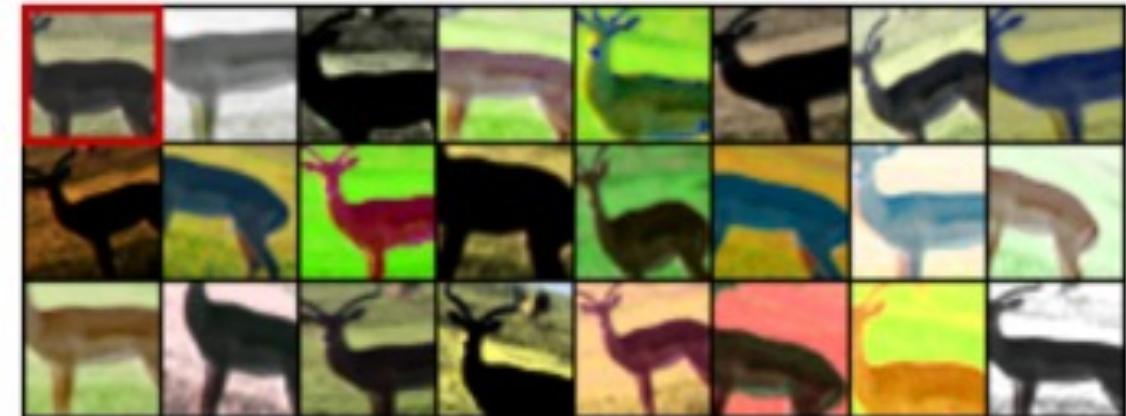


Fig. 2. Several random transformations applied to one of the patches extracted from the STL unlabeled dataset. The original ('seed') patch is in the top left corner.

# Подходы в самообучении

## Решение головоломки

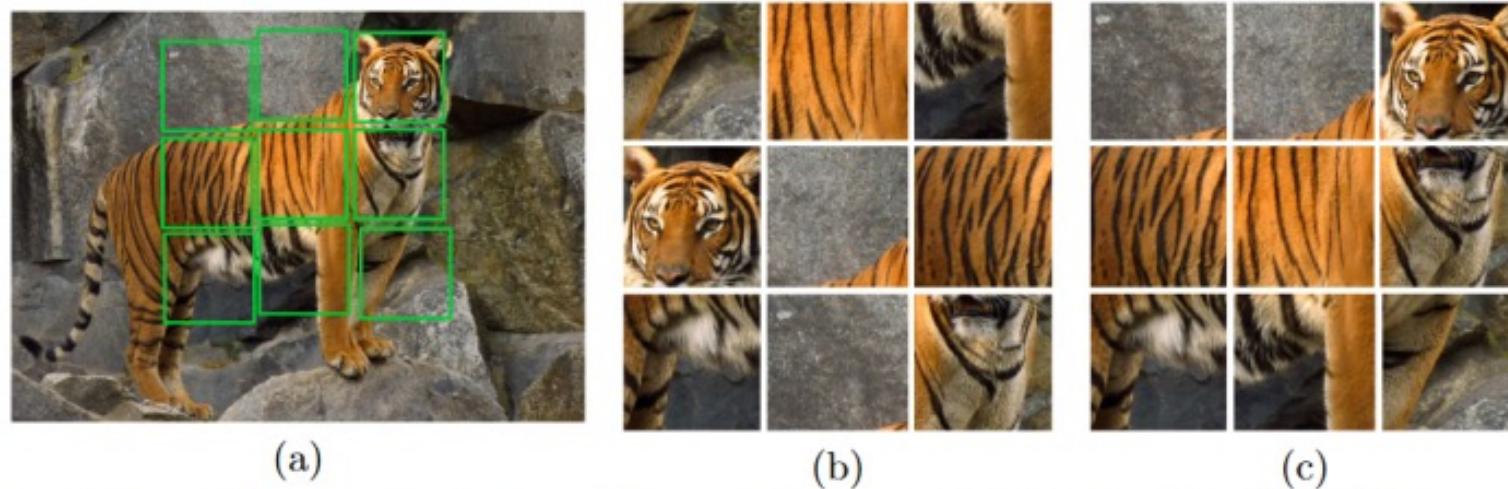


Fig. 1: Learning image representations by solving Jigsaw puzzles. (a) The image from which the tiles (marked with green lines) are extracted. (b) A puzzle obtained by shuffling the tiles. Some tiles might be directly identifiable as object parts, but others are ambiguous (*e.g.*, have similar patterns) and their identification is much more reliable when all tiles are jointly evaluated. In contrast, with reference to (c), determining the relative position between the central tile and the top two tiles from the left can be very challenging [10].

# Подходы в самообучении

## Решение головоломки

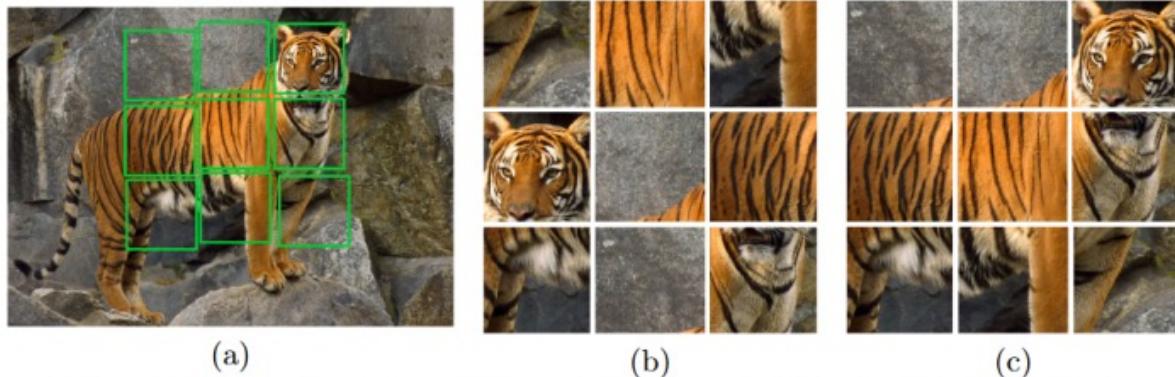


Fig. 1: Learning image representations by solving Jigsaw puzzles. (a) The image from which the tiles (marked with green lines) are extracted. (b) A puzzle obtained by shuffling the tiles. Some tiles might be directly identifiable as object parts, but others are ambiguous (*e.g.*, have similar patterns) and their identification is much more reliable when all tiles are jointly evaluated. In contrast, with reference to (c), determining the relative position between the central tile and the top two tiles from the left can be very challenging [10].

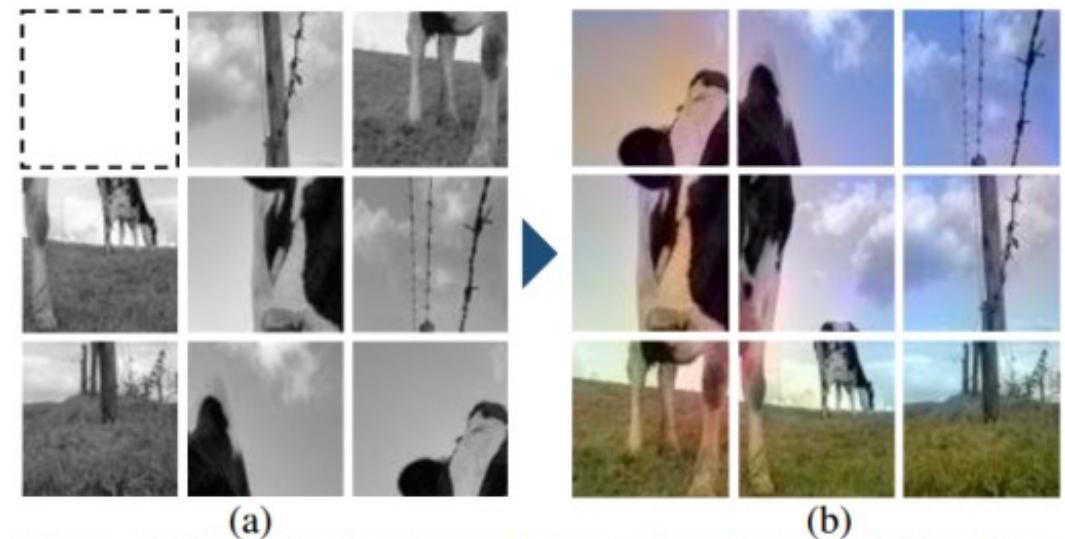


Figure 1. **Learning image representations by completing damaged jigsaw puzzles.** We sample 3-by-3 patches from an image and create damaged jigsaw puzzles. (a) is the puzzles after shuffling the patches, removing one patch, and decolorizing. We push a network to recover the original arrangement, the missing patch, and the color of the puzzles. (b) shows the outputs; while the pixel-level predictions are in *ab* channels, we visualize with their original *L* channels for the benefit of the reader.

\*Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In European Conference on Computer Vision (ECCV), 2016.

\*\* Kim, D., Cho, D., Yoo, D., and Kweon, I. S. Learning Image Representations by Completing Damaged Jigsaw Puzzles. In WACV 2018, 2018.

# Подходы в самообучении

## Контекстные кодировщики

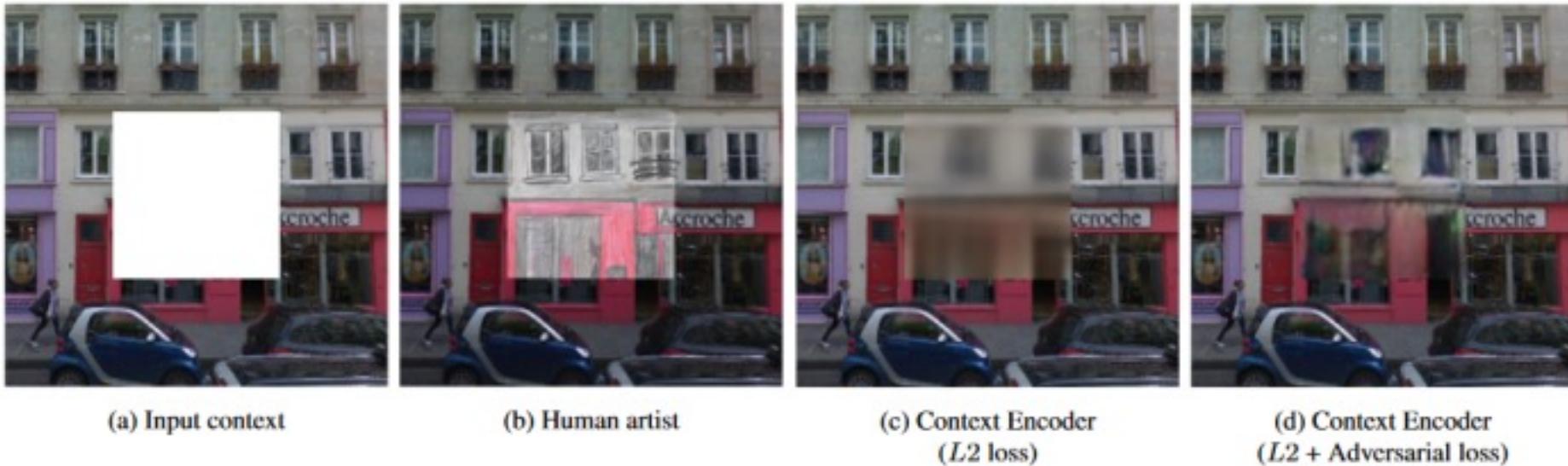


Figure 1: Qualitative illustration of the task. Given an image with a missing region (a), a human artist has no trouble inpainting it (b). Automatic inpainting using our *context encoder* trained with  $L_2$  reconstruction loss is shown in (c), and using both  $L_2$  and adversarial losses in (d).

# Задача верификации

Постановка задачи – на заданных двух изображениях один и тот же объект?

- Алгоритм должен иметь класс «unseen»
- Нерепрезентативная выборка (мало данных, дисбаланс классов)

Face recognition dataset (MSRA-CF)

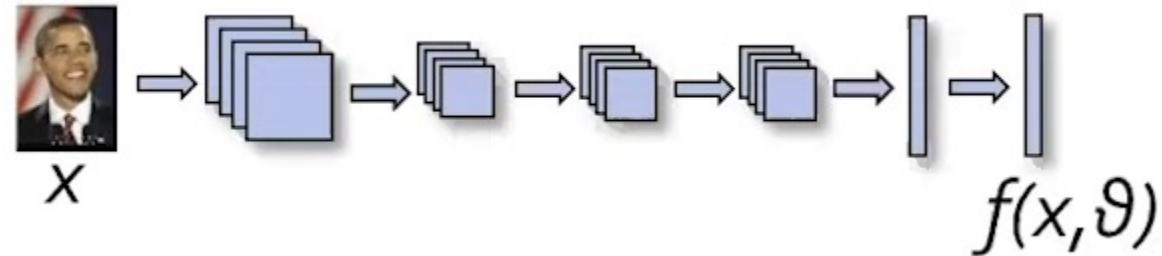


Re-identification dataset (ViPER)

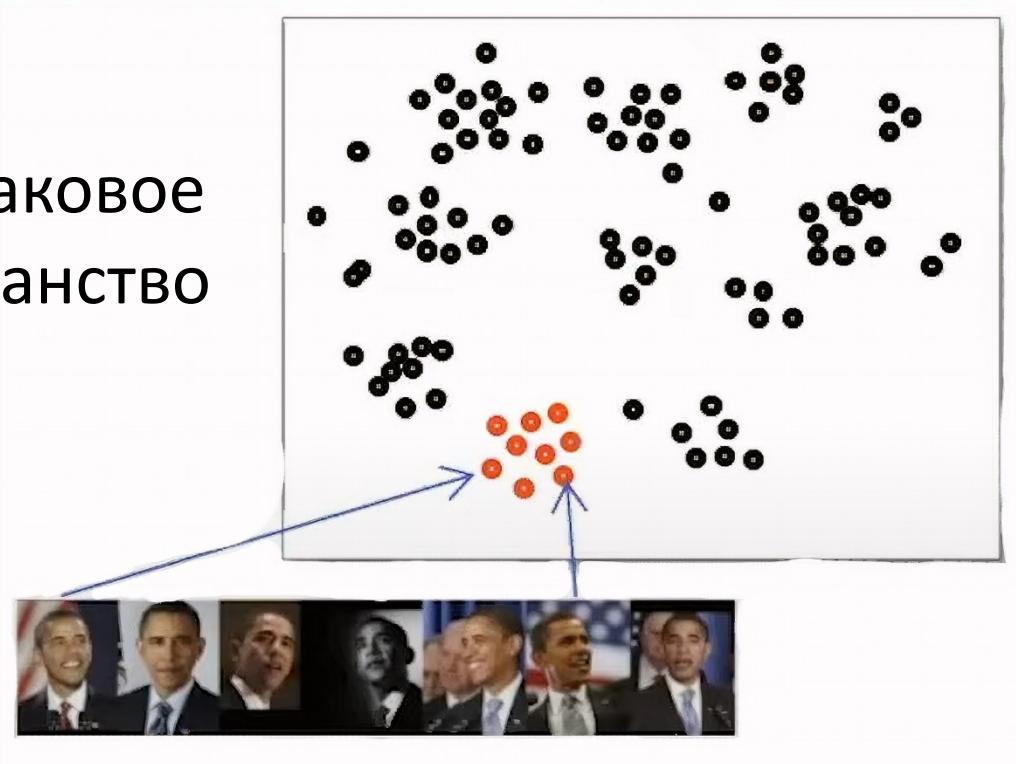


Open world dataset

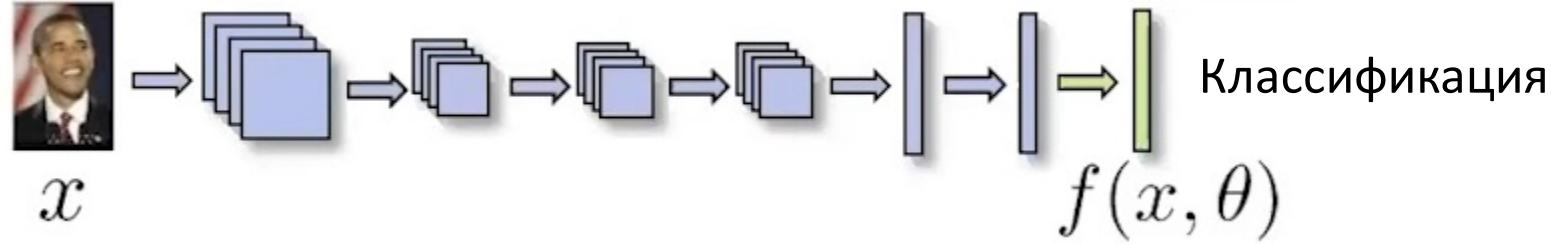
# Задача верификации как embedding learning



Признаковое  
пространство

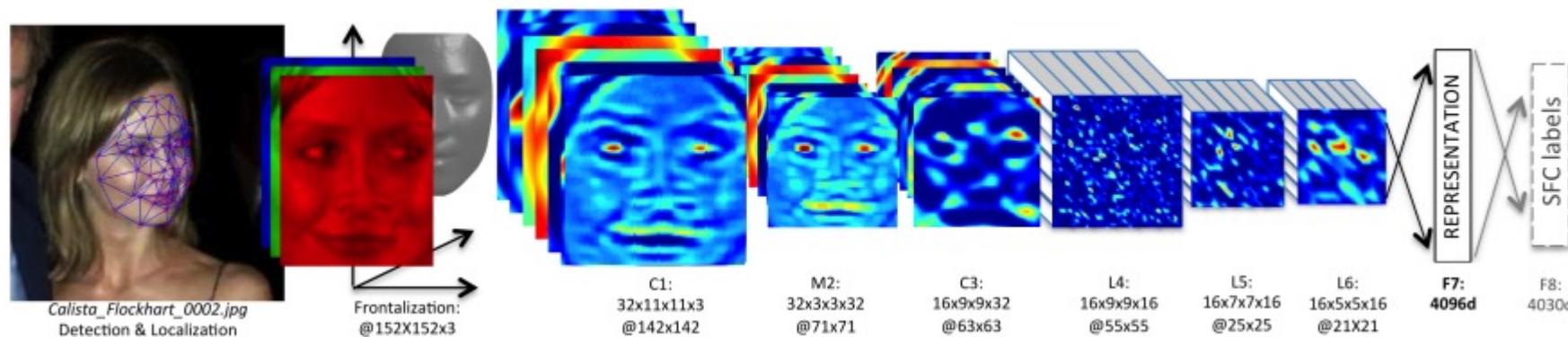


# Задача верификации как embedding learning

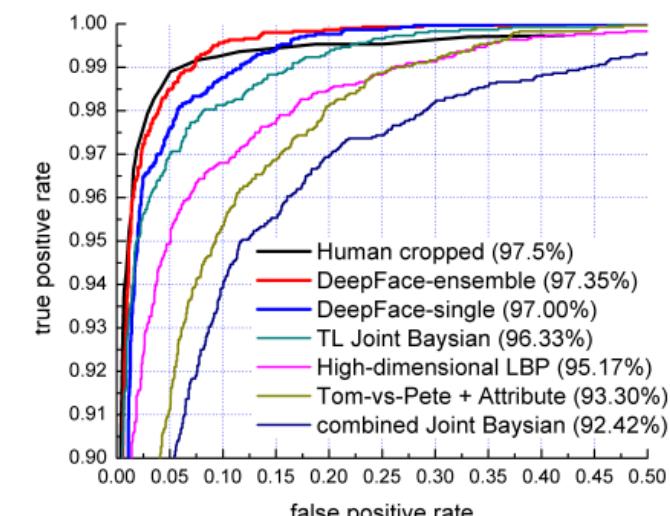


- Обучение на задаче классификации
- Чем больше данных, тем лучше результат
- Классы во время обучения могут быть в виде прототипов классов на тестировании

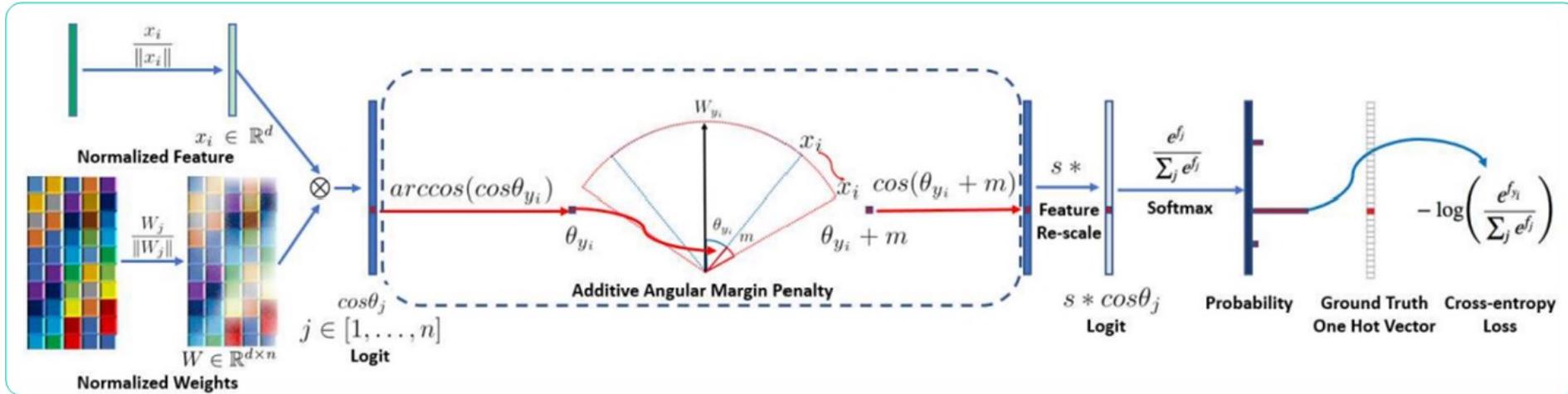
# Верификация: «DeepFace»



- Classification network обучена на 4030 человек (класс) ~ 1000 изображений
- Постановка задачи: один vs другие

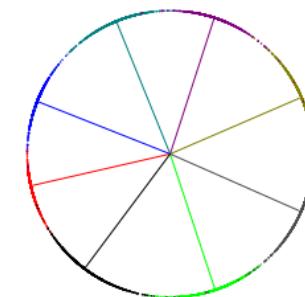


# Идея нормализации и отступа

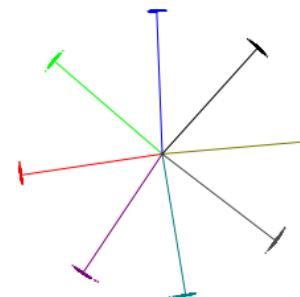


Loss Functions	LFW	CFP-FP	AgeDB-30
ArcFace (0.4)	99.53	95.41	94.98
ArcFace (0.45)	99.46	95.47	94.93
<b>ArcFace (0.5)</b>	<b>99.53</b>	<b>95.56</b>	<b>95.15</b>
ArcFace (0.55)	99.41	95.32	95.05
SphereFace [18]	99.42	-	-
SphereFace (1.35)	99.11	94.38	91.70
CosFace [37]	99.33	-	-
CosFace (0.35)	99.51	95.44	94.56
CM1 (1, 0.3, 0.2)	99.48	95.12	94.38
CM2 (0.9, 0.4, 0.15)	99.50	95.24	94.86
Softmax	99.08	94.39	92.33
Norm-Softmax (NS)	98.56	89.79	88.72

$$\text{ArcFace loss: } L_3 = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{s(\cos(\theta_{y_i} + m))}}{e^{s(\cos(\theta_{y_i} + m))} + \sum_{j=1, j \neq y_i}^n e^{s \cos \theta_j}}$$

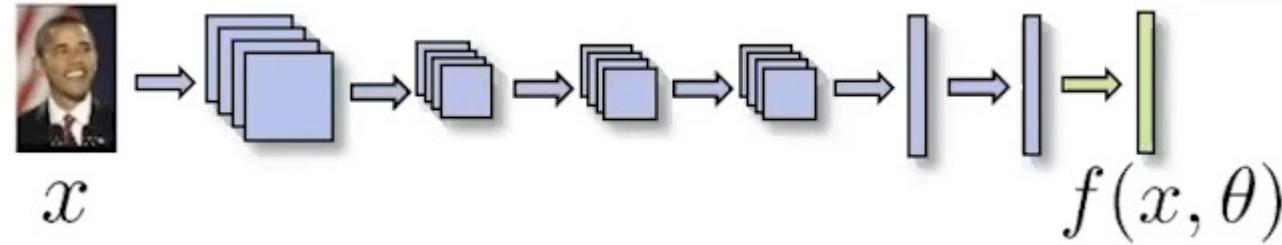


(a) Softmax



(b) ArcFace

# Pair-based learning (contrastive)



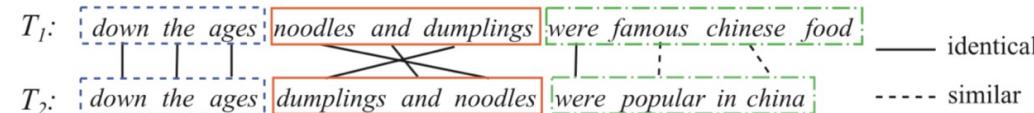
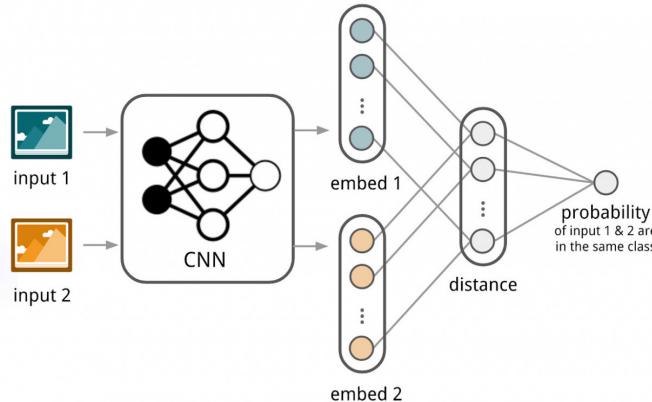
$$L^+((x_1, x_2); \theta) = \rho(f(x_1, \theta), f(x_2, \theta))$$

$$L^-((x_1, x_2); \theta) = \max(0, M - \rho(f(x_1, \theta), f(x_2, \theta)))$$

Функции расстояний:

- $1 - \cos$
- L2 (batch norm)
- Separate network

# FaceNet



**Примитивный triplet loss:** 
$$\sum_i^N \left[ \|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+$$

- Сиамские нейронные сети
- Использовать большие партии (mini-batches) – 1800, 40 изображений для разных классов + рандом
- Взять все positives для всей партии (batch)
- Взять «semi-hard» negatives:

$$\|f(x_i^a) - f(x_i^p)\|_2^2 < \|f(x_i^a) - f(x_i^n)\|_2^2$$

# FaceNet: результаты



- Результаты 99.63% на LFW  
(точность человека ~97%)

Точность и размер  
выборка

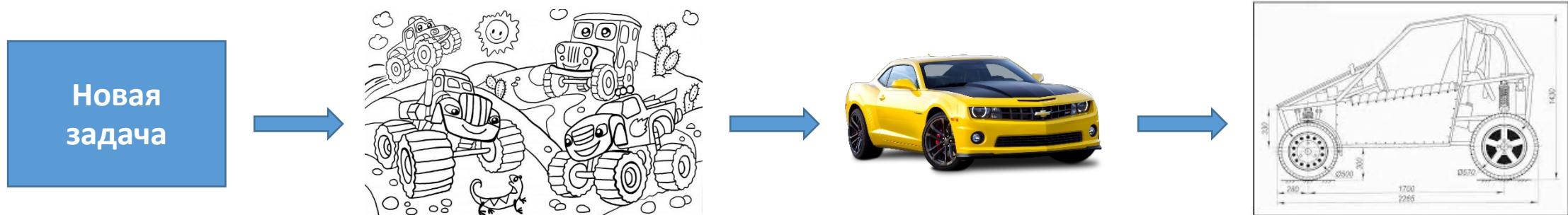
#training images	VAL
2,600,000	76.3%
26,000,000	85.1%
52,000,000	85.1%
260,000,000	86.2%

# Self-supervised feature learning

Общий подход:

- Взять модель, предобученную на текущих неразмеченных данных
- Дообучение (fine-tune) на новую задачу на размеченных данных

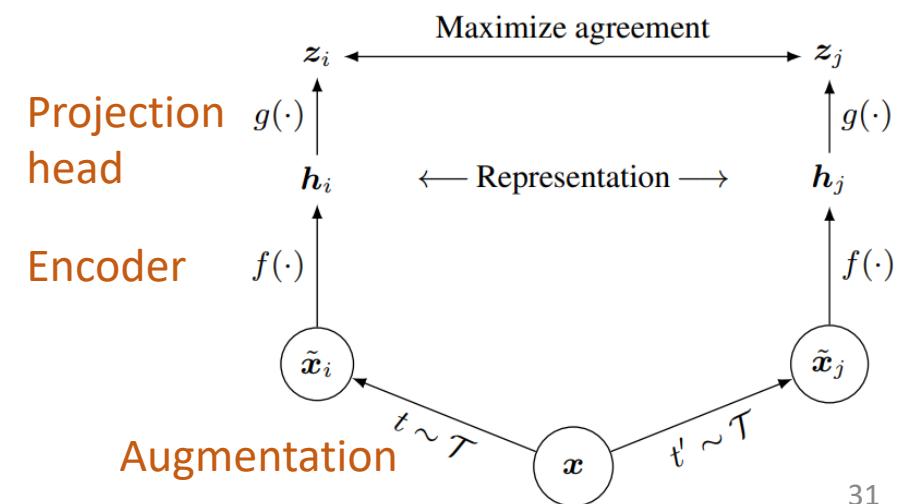
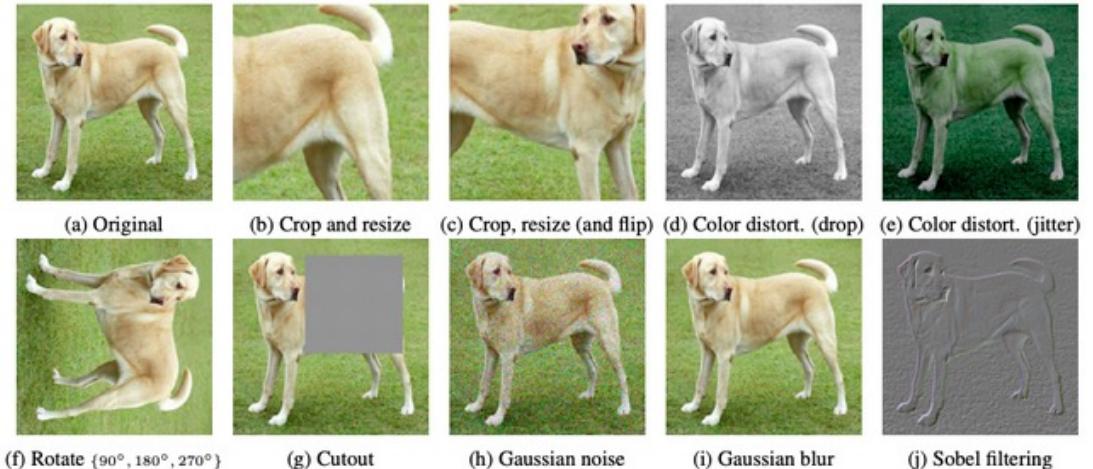
Альтернатива – взять предобученную модель на ImageNet



# SimCLR

- Каждый batch содержит пары изображений
- Каждая пара является измененными версиями исходного изображения из выборки с изменениями типа crop, Gaussian blur, color distortion
- Цель – обучить модель сопоставлять пары, классифицировать как один объект (класс)
- Требуется большой размер пакета (batch)
- Loss:

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)}$$



# Приложение

(1) Self-supervised learning on **unlabeled** natural images



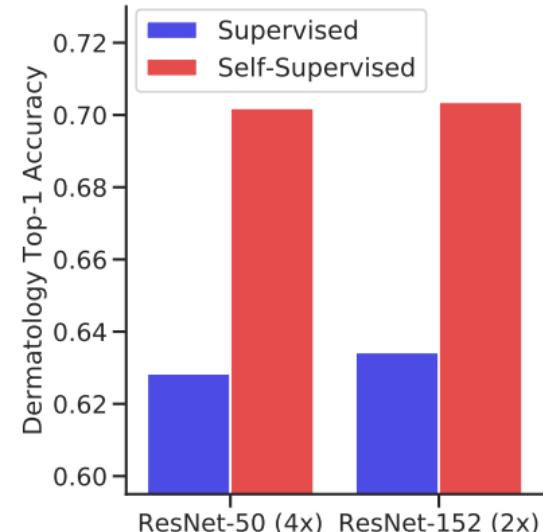
(2) Self-supervised learning on **unlabeled** medical images and **Multi-Instance Contrastive Learning (MICLe)** if multiple images of each medical condition are available



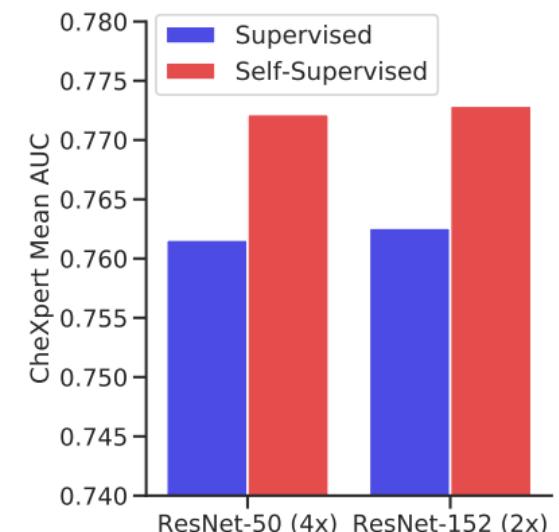
(3) Supervised fine-tuning on **labeled** medical images



15 000 обучающих  
примеров,  
454 000  
неразмеченных,  
27 классов



67 000 обучающих  
примеров,  
112 000  
неразмеченных,  
5 классов



Синее – предобучено на ImageNet  
Красное – unsupervised ImageNet + Target domain

# Заключение

- Рассмотрены задачи технического зрения
- Обзор технологии переноса обучения и самообучения
- Задача верификации - Embedding learning

# Ссылки

Kaggle - [Humpback Whale Identification](#), пример на задачу  
верификации

Kaggle - [Deepfake Detection Challenge](#), пример анти-фрод  
систем

Kaggle - [Google Landmark Recognition 2020](#), пример задачи  
на открытый класс