Лаборатория компьютерной
графики и мультимедиа
ВМК МГУ имени М.В. Ломоносова

Курс «Компьютерное зрение»

«Само-обучение и фундаментальные модели»

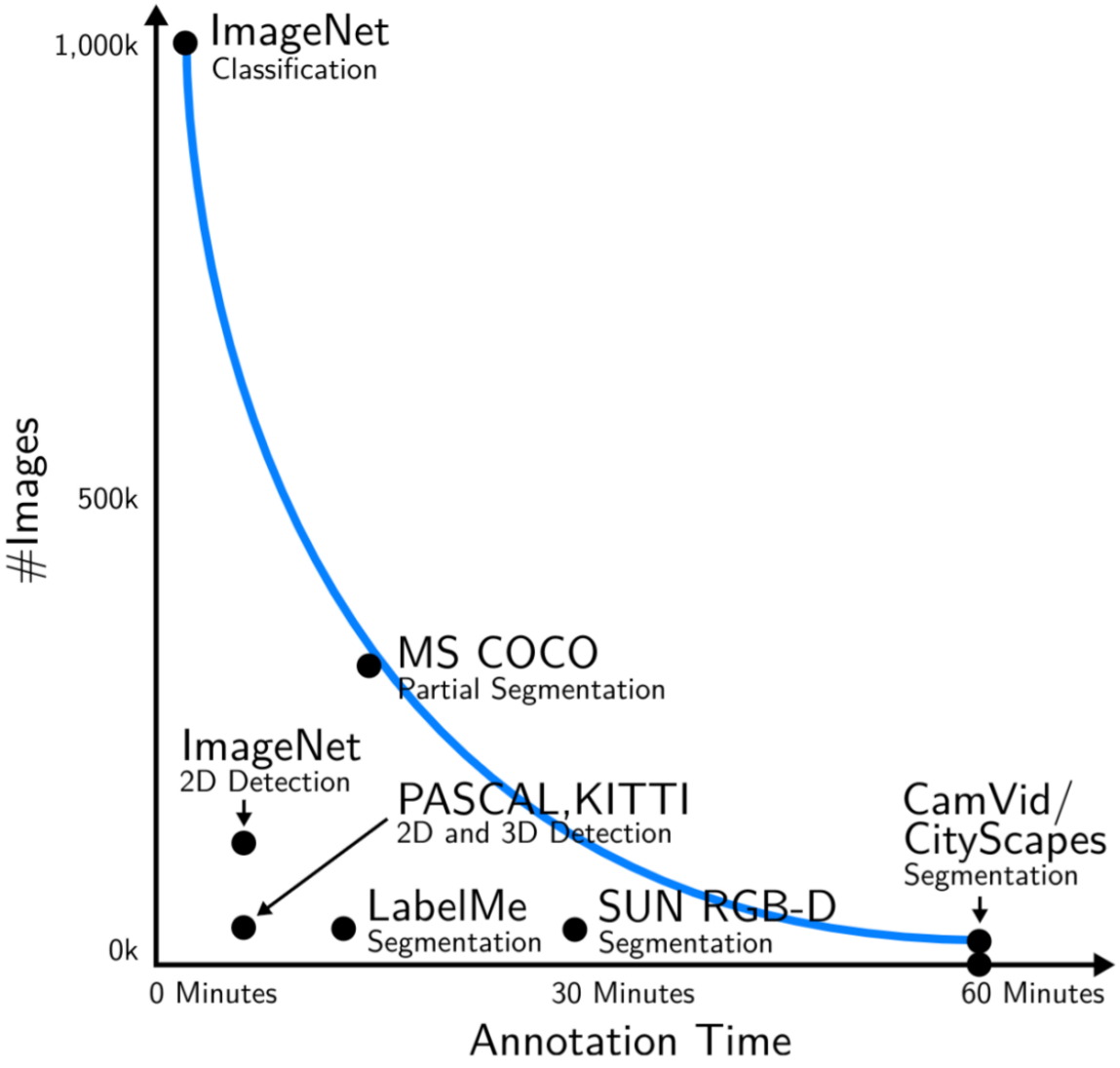
Антон Конушин и Тимур Мамедов

2025 год

План лекции

1. Введение
2. Прокси-задачи
3. Констранстное обучение и маскирование
4. Фундаментальные модели

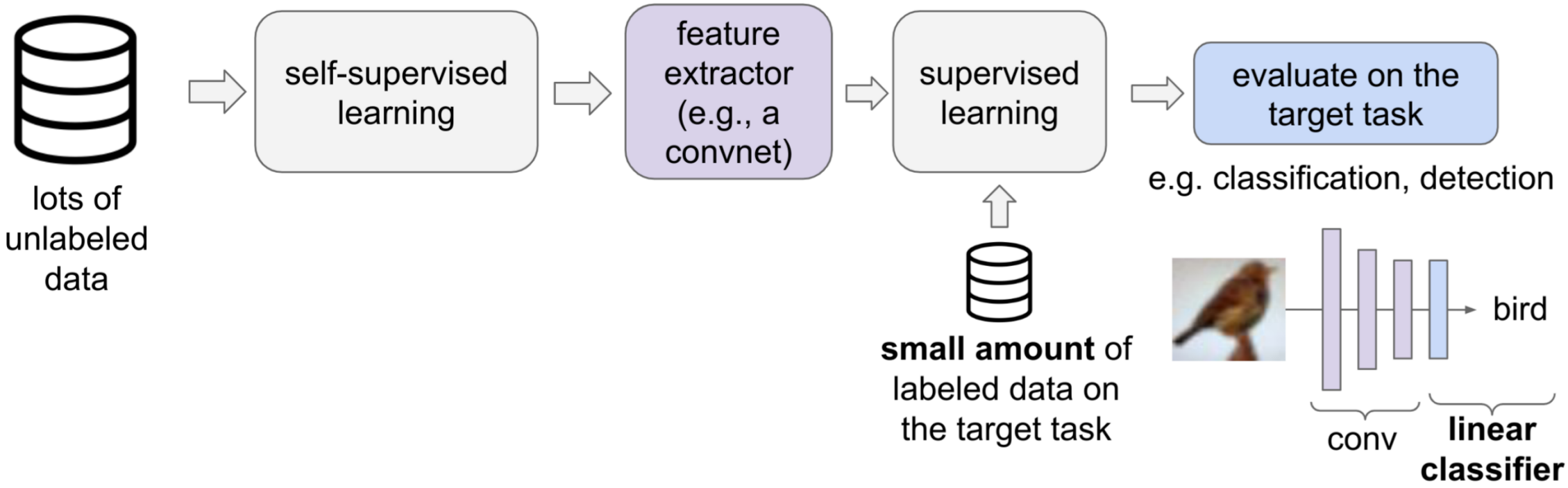
Дополнительные plug-in модули



- Разметка данных дорогая
- Метки всегда содержат ошибки
- Обучать людей размечать эффективно и без ошибок очень сложно
- Сушествует много неразмеченных данных, м.б. мы можем как-то их использовать?

Xie et al. Semantic Instance Annotation of Street Scenes by 3D to 2D Label Transfer. CVPR 2016

Предобучение через само-обучение

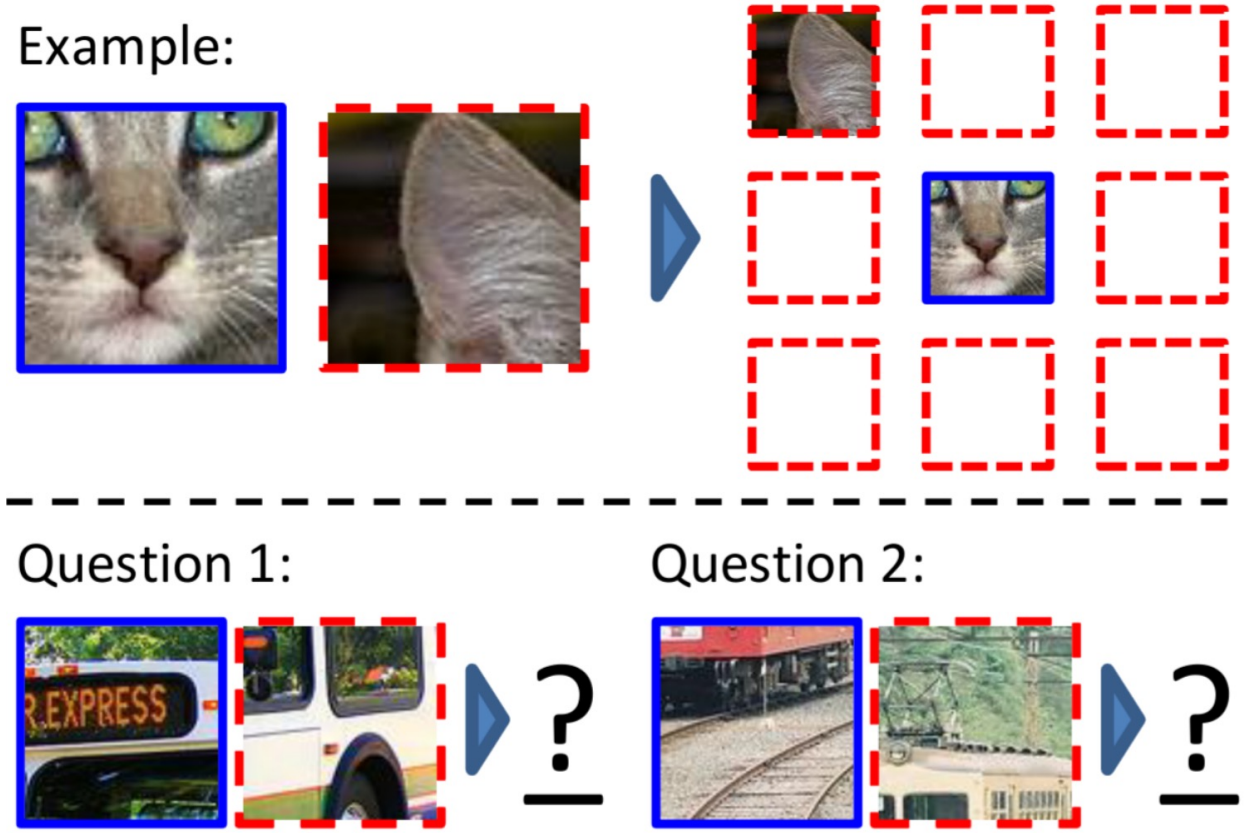


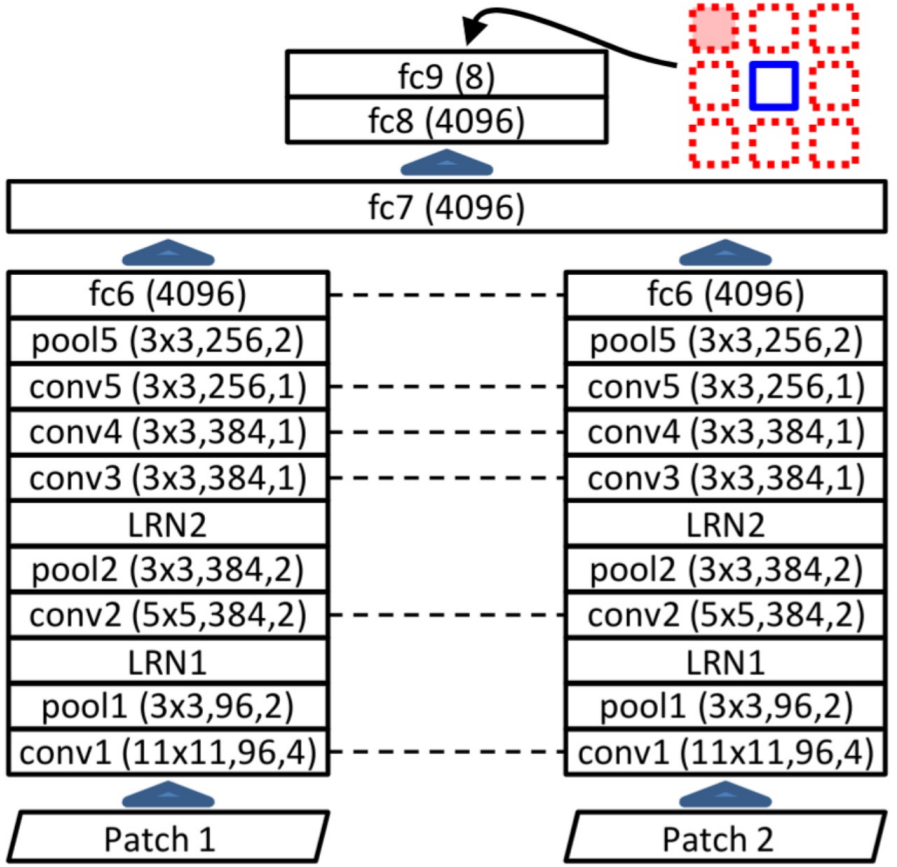
1. Обучить полезный нейро-признаки (или представления / representation) с помощью прокси-задачи
2. Добавить поверх нейропризнаков сколько-то слоёв и до-обучить на размеченных датасетах

План лекции

1. Введение
2. Прокси-задачи
3. Констранстное обучение и маскирование
4. Фундаментальные модели

Предсказание контекста

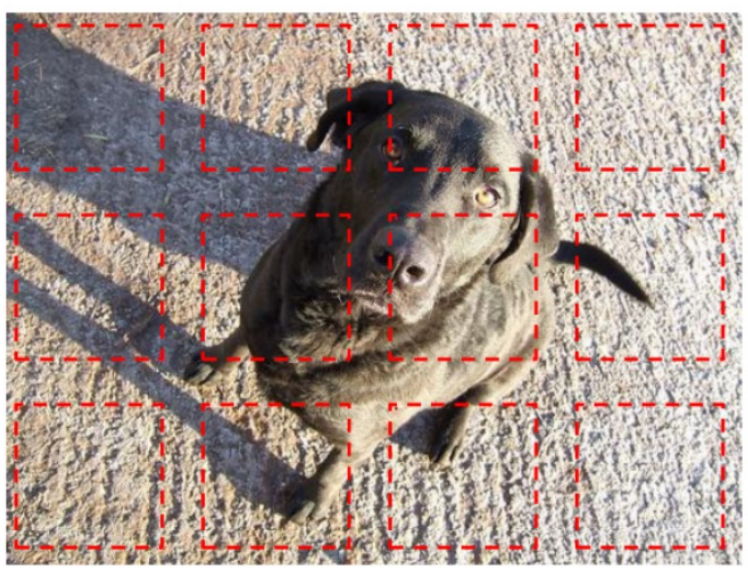




Предсказываем, где расположен фрагмент относительно центрального, как задача 8-и классовой классификаци

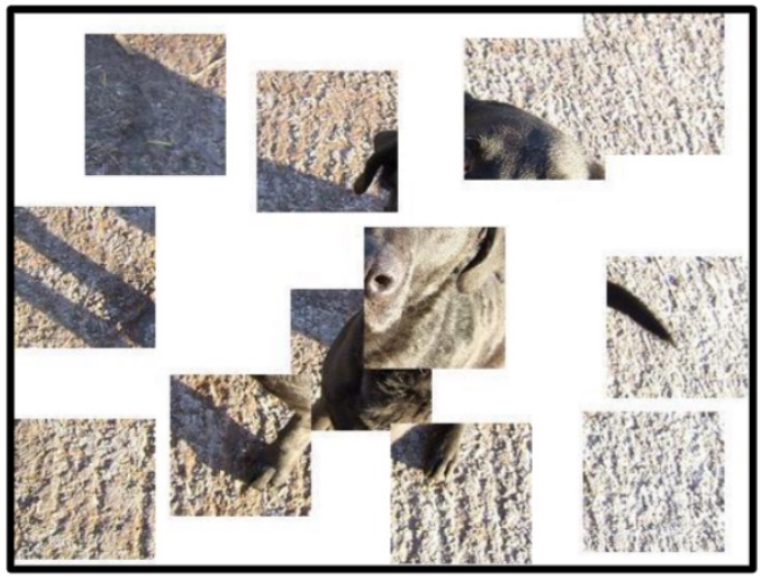
Doersch et al. Unsupervised Visual Representation Learning by Context Prediction. ICCV 2015

Предсказание контекста – забавный эффект

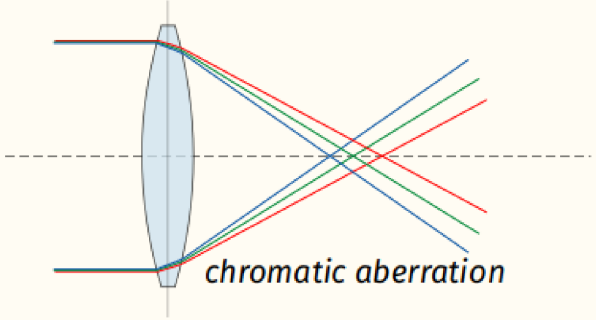


Initial layout, with sampled patches in red

Image layout is discarded



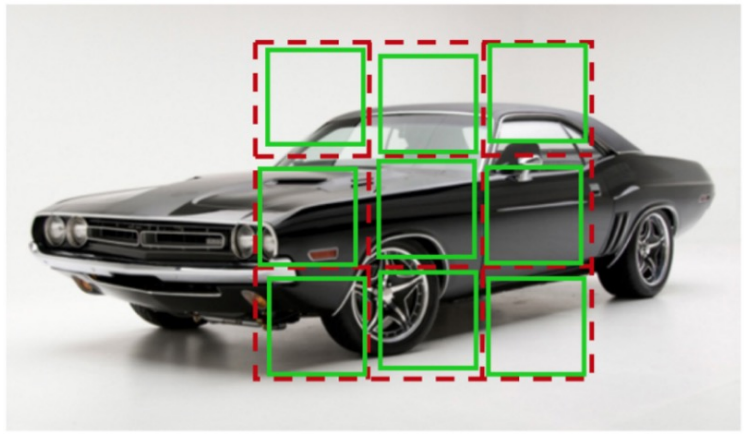
We can recover image layout automatically



Сеть может научиться «жульничать», предсказывая положения патча по эффекту хроматической абберации

Doersch et al. Unsupervised Visual Representation Learning by Context Prediction. ICCV 2015

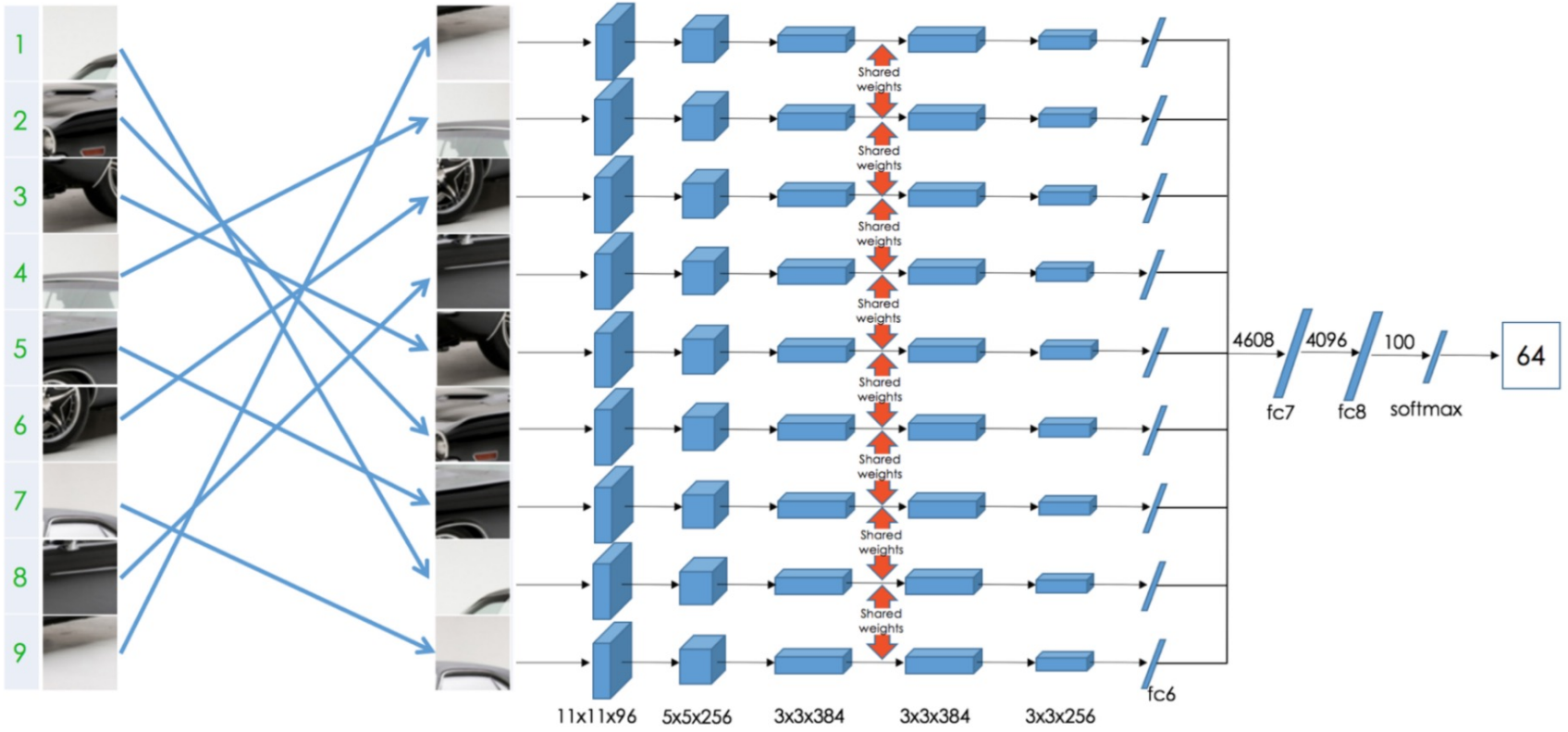
Jigsaw puzzle



Permutation Set

index	permutation
64	9,4,6,8,3,2,5,1,7

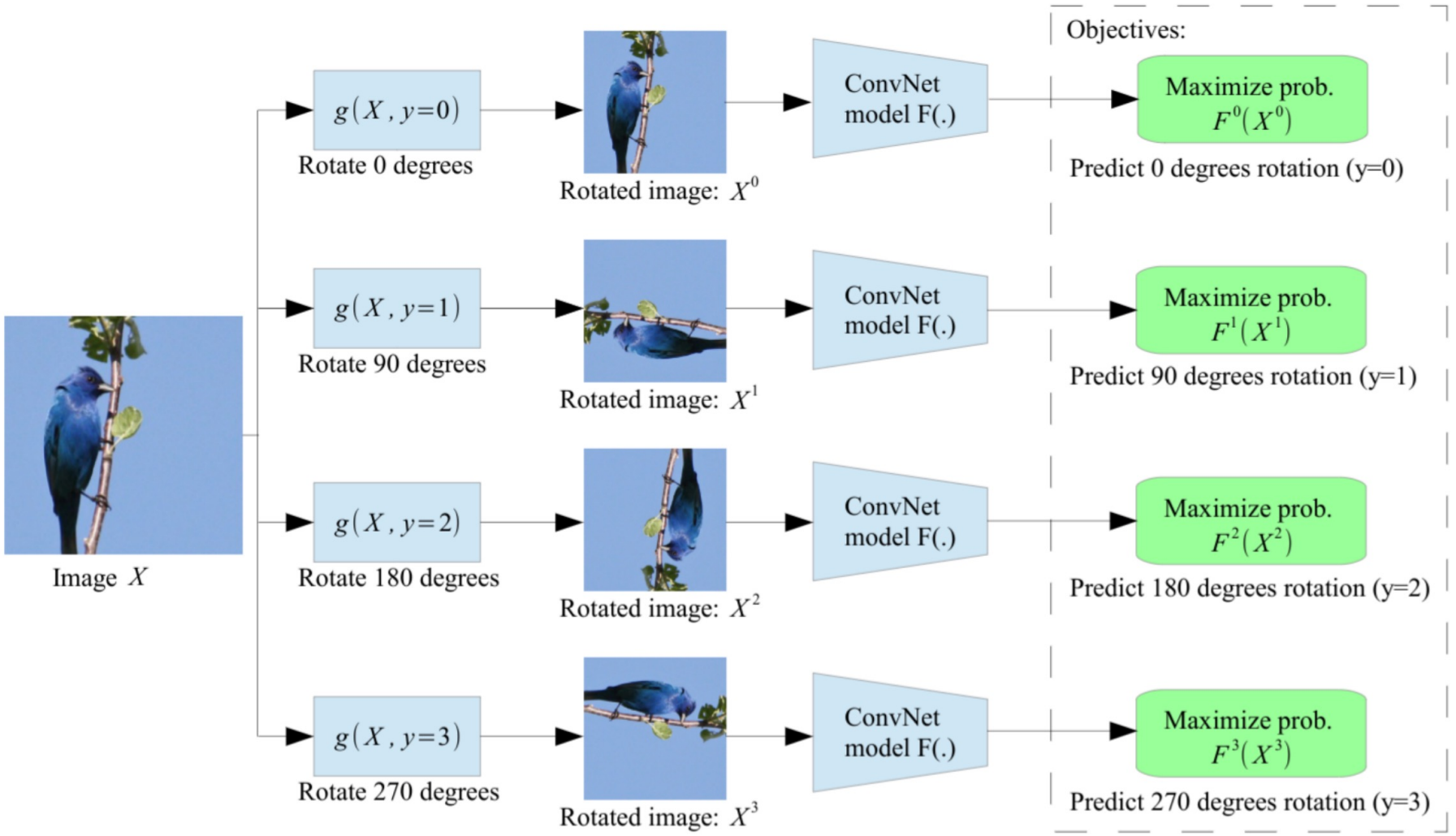
Reorder patches according to the selected permutation



Предсказываем «перестановку» патчей – как их нужно переставить, чтобы разместить фрагменты в правильном порядке (~1000 классов)

Noroozi, Favaro. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. ECCV 2016

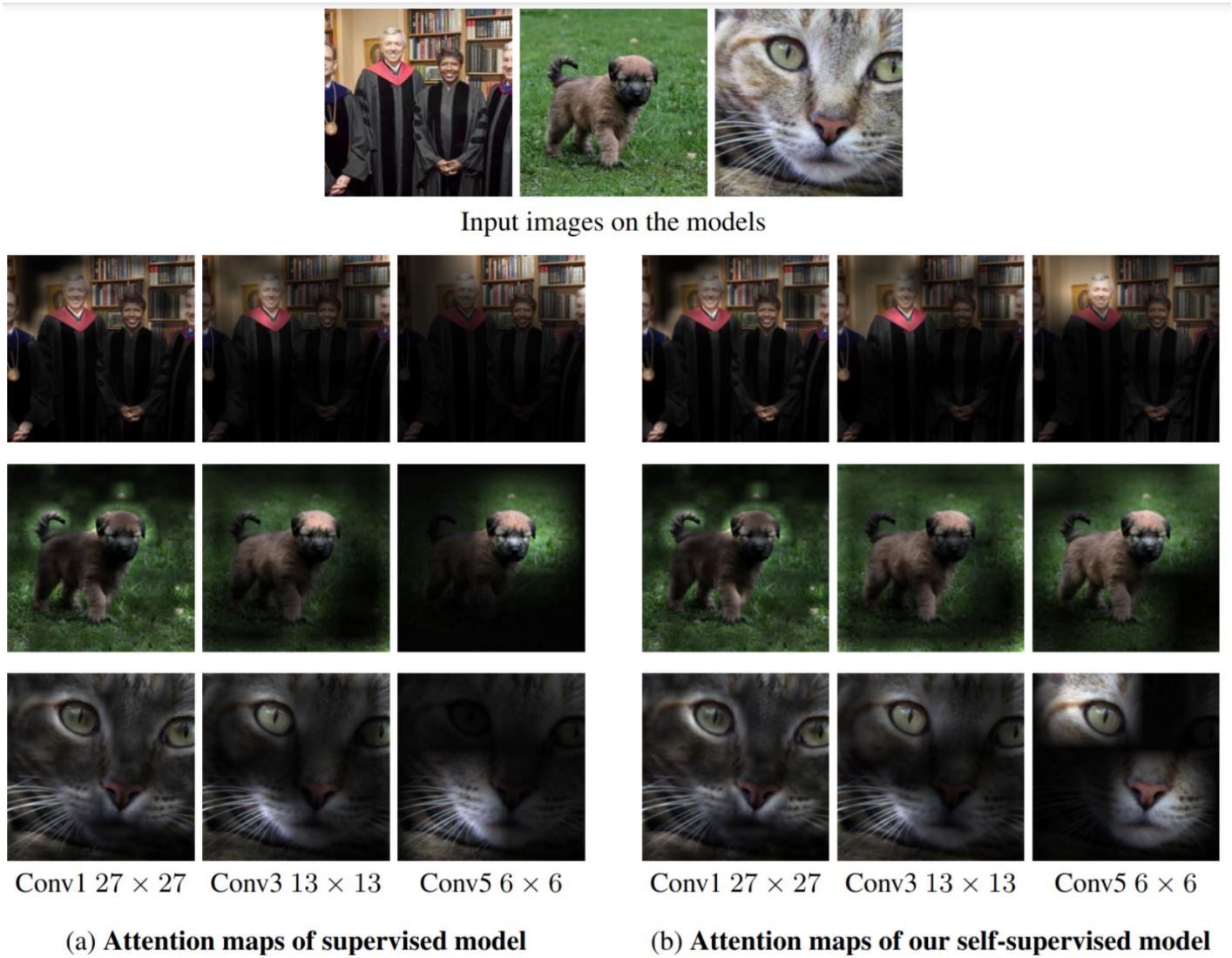
Повороты



Предсказываем поворот фрагмента относительно базового

Gidaris et al. Unsupervised Representation Learning by Predicting Image Rotations. ICLR 2018

Повороты



Модель учится фокусироваться на ключевых фрагментах изображений

Gidaris et al. Unsupervised Representation Learning by Predicting Image Rotations. ICLR 2018

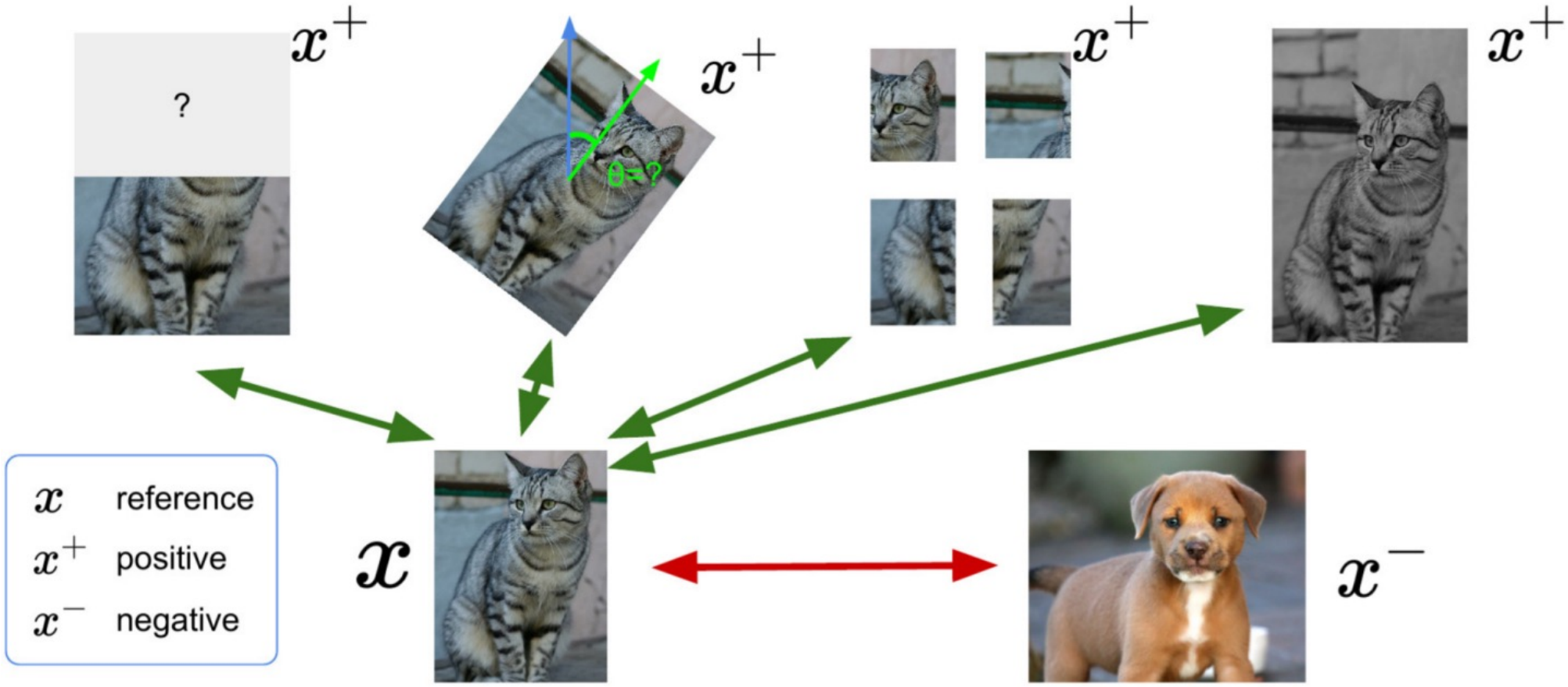
Резюме прокси-задач

- Прокси-задачи строятся эмпирически, опираясь на какую-то априорную визуальную информацию, например, предсказание поворотов, положения в пространстве, цветов изображений
- Модели учатся понимать суть изображения для решения прокси-задачи
- Полезность обученных признаков оценивается на целевой задачей
- Придумать прокси-задачу сложно
- Сложно предсказать, насколько полезны обученные признаки и насколько модель хорошо обобщается

План лекции

1. Введение
2. Прокси-задачи
3. Констранстное обучение и маскирование
4. Фундаментальные модели

Контрастное обучение (Contrastive learning)



Given a score function $s(\cdot, \cdot)$, we want to learn a mapping f_θ that yields high score for positive pairs (x, x^+) and low score for negative pairs (x, x^-):

$$s(f_\theta(x), f_\theta(x^+)) \gg s(f_\theta(x), f_\theta(x^-))$$

Идея контрастного обучения

Пусть у нас 1 референсное изображений (x), 1 положительный пример (x^+) и $N-1$ отрицательных примеров (x^-). Будем использовать многоклассовую кросс-энтропию:

$$\mathscr{L} = -\mathbf{E}_{\mathscr{X}}\left[\log \frac{\exp\left(s(f_\theta(x), f_\theta(x^+))\right)}{\exp\left(s(f_\theta(x), f_\theta(x^+))\right) + \sum_{j=1}^{N-1}\exp\left(s(f_\theta(x), f_\theta(x_j^-))\right)}\right]$$

Функция известна как InfoNCE loss и её отрицание дает нижнюю границу на mutual information между $\mathrm{f}_\theta(\mathrm{x})$ и $\mathrm{f}_\theta(\mathrm{x}^+)$:

$$MI[f_\theta(x), f_\theta(x^+)] \geq \log N - \mathscr{L}$$

Максимизация совместной информации между разными «видами» на изображение позволяет извлечь высокуровненывую информацию из изображений

Oord et al. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748

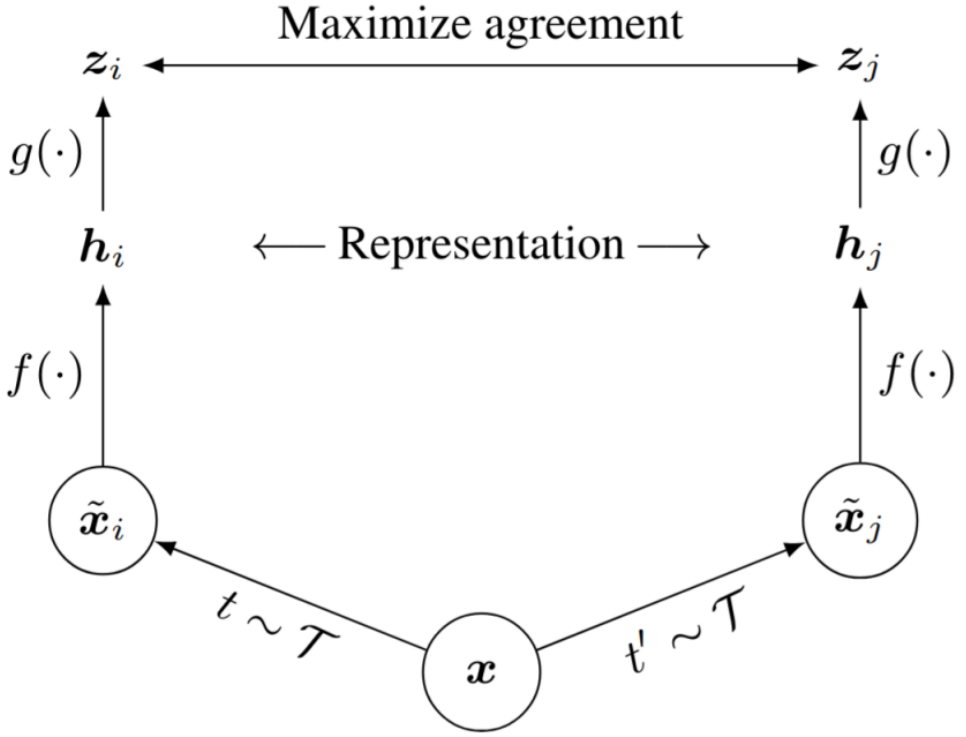
SimCLR

Функция косинусного расстояния как целевая функция:

$$s(\boldsymbol{z}_i, \boldsymbol{z}_j) = \frac{\boldsymbol{z}_i^T \boldsymbol{z}_j}{\|\boldsymbol{z}_i\| \, \|\boldsymbol{z}_j\|}$$

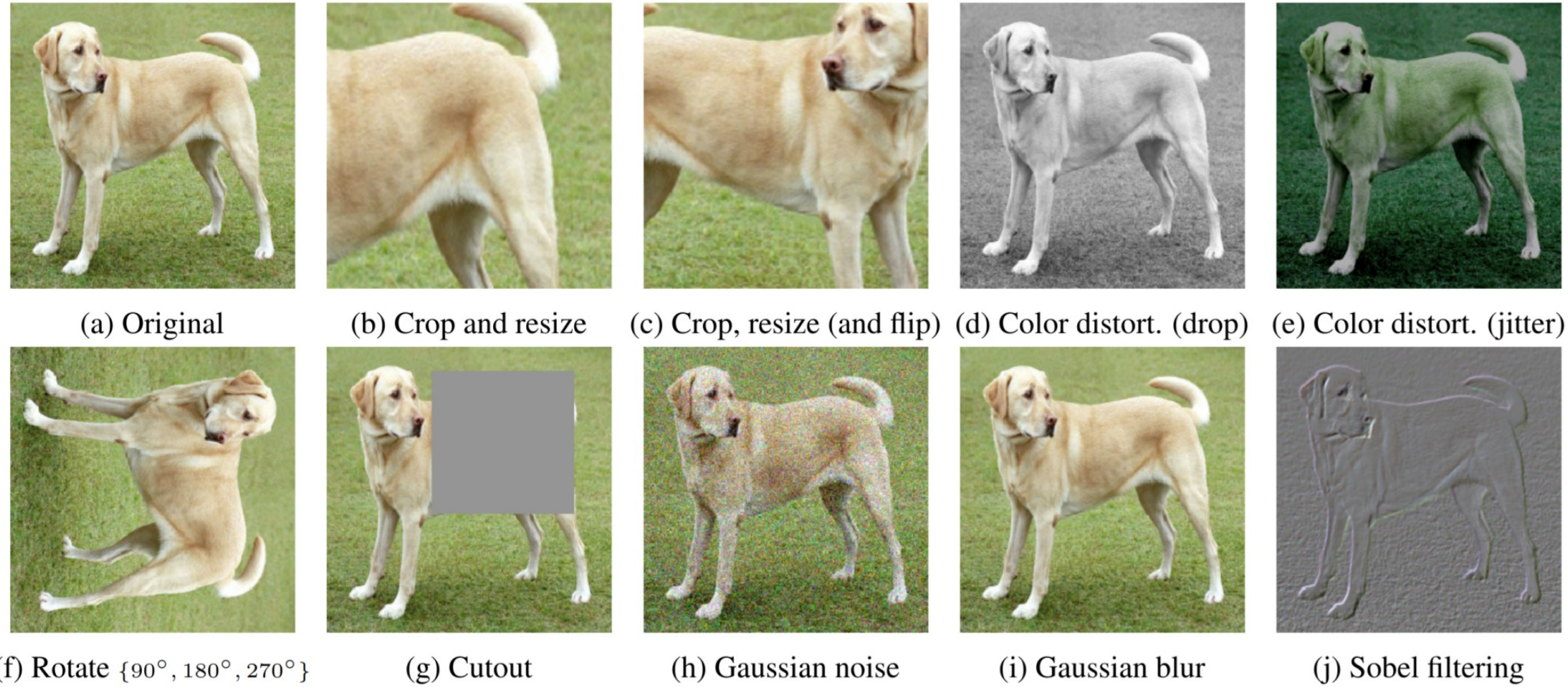
Используемся вспомогательную сеть-проектор g(.) для проецирования признаков в пространство, в котором применяем контрастное обучение

Такой проектор улучшает обучение, т.к. более важная информация сохраняется в h



Chen et al. A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020

SimCLR



Примеры аугментаций изображений при получении пар изображений

Chen et al. A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020

SimCLR

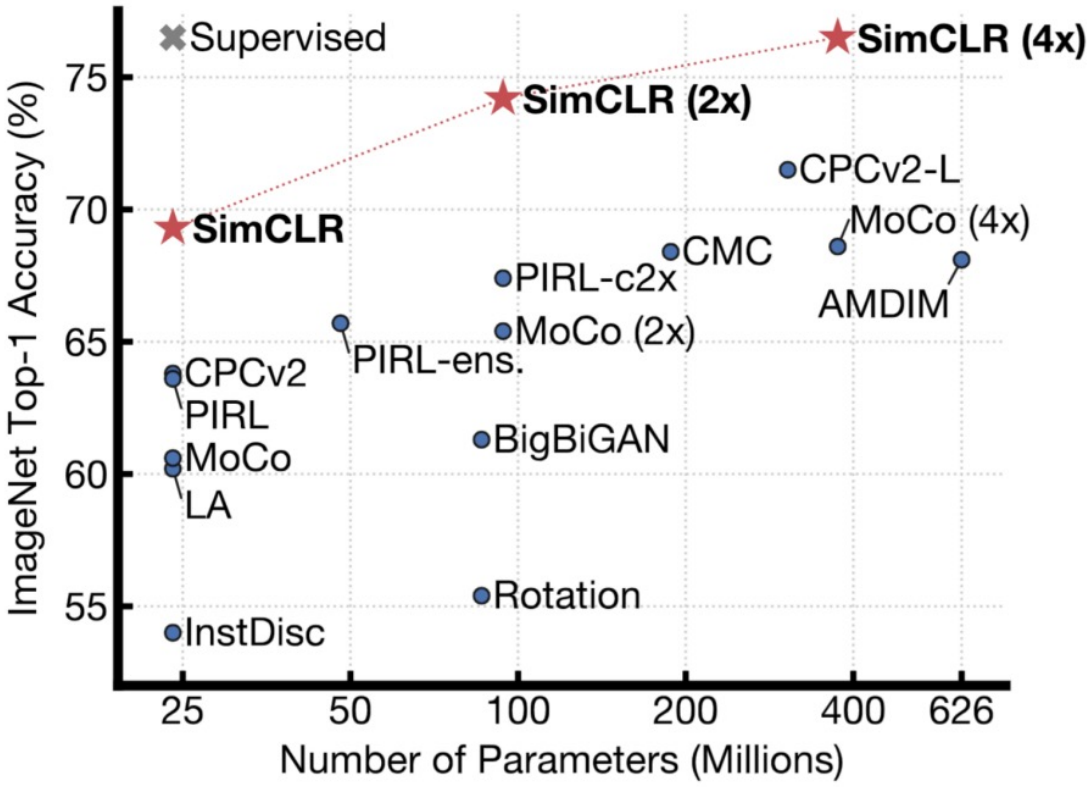
Algorithm 1 SimCLR's main learning algorithm.

input: batch size N, constant τ, structure of $f, g, \mathcal{T}$.
for sampled minibatch $\{\boldsymbol{x}_k\}_{k=1}^{N}$ **do**
  for all $k \in \{1, \ldots, N\}$ **do**
    draw two augmentation functions $t \sim \mathcal{T}, t' \sim \mathcal{T}$
    # the first augmentation
    $\tilde{\boldsymbol{x}}_{2k-1} = t(\boldsymbol{x}_k)$
    $\boldsymbol{h}_{2k-1} = f(\tilde{\boldsymbol{x}}_{2k-1})$ # representation
    $\boldsymbol{z}_{2k-1} = g(\boldsymbol{h}_{2k-1})$ # projection
    # the second augmentation
    $\tilde{\boldsymbol{x}}_{2k} = t'(\boldsymbol{x}_k)$
    $\boldsymbol{h}_{2k} = f(\tilde{\boldsymbol{x}}_{2k})$ # representation
    $\boldsymbol{z}_{2k} = g(\boldsymbol{h}_{2k})$ # projection
  end for
  for all $i \in \{1, \ldots, 2N\}$ and $j \in \{1, \ldots, 2N\}$ **do**
    $s_{i,j} = \boldsymbol{z}_i^\top \boldsymbol{z}_j / (\|\boldsymbol{z}_i\| \|\boldsymbol{z}_j\|)$ # pairwise similarity
  end for
  define $\ell(i, j)$ **as** $\ell(i, j) = -\log \frac{\exp(s_{i,j}/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(s_{i,k}/\tau)}$
  $\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} [\ell(2k-1, 2k) + \ell(2k, 2k-1)]$
  update networks f and g to minimize $\mathcal{L}$
end for
return encoder network $f(\cdot)$, and throw away $g(\cdot)$

Для каждого примера в минибатче сэплируем две функции аугментации, применяем их, и считаем расстояние между признаками

Chen et al. A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020

SimCLR

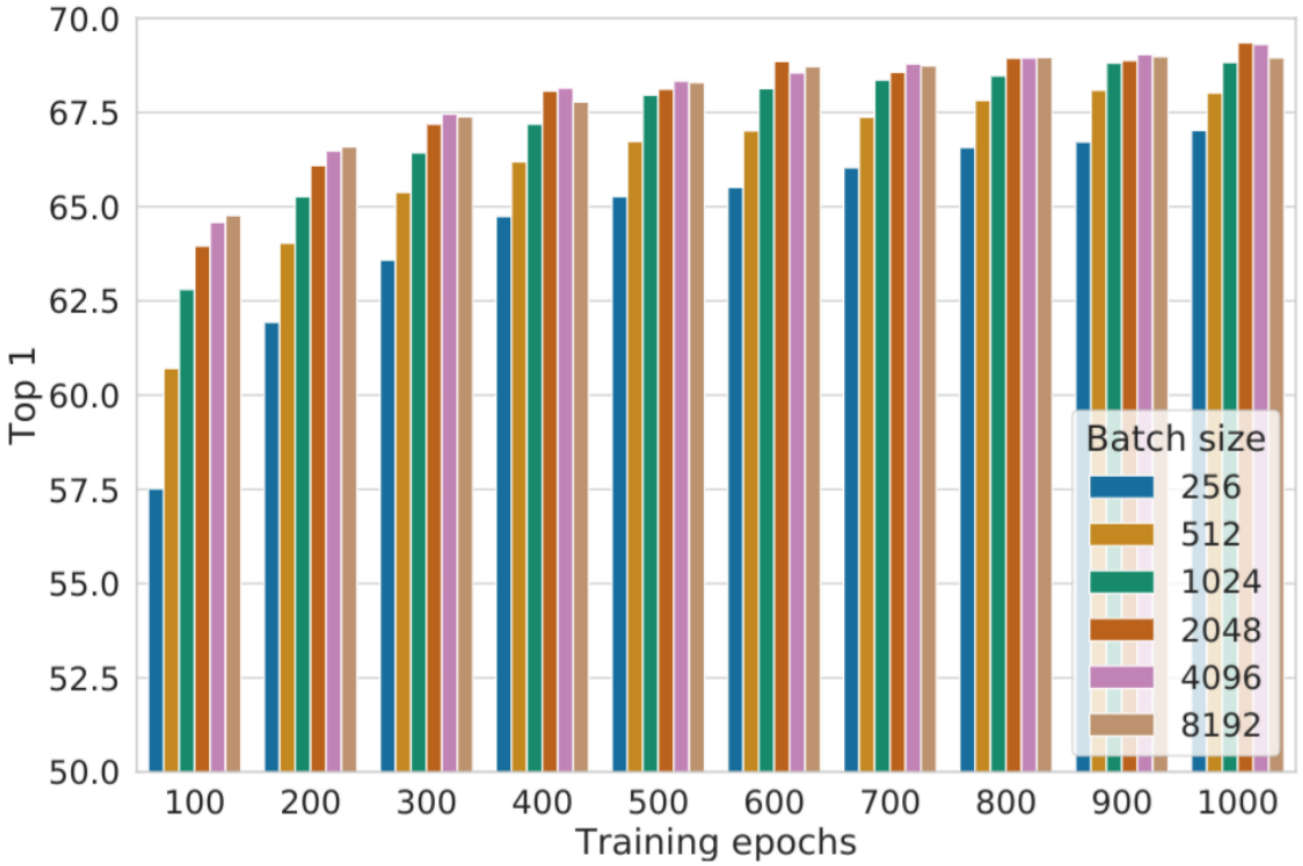


Method	Architecture	Label fraction 1%	Label fraction 10%
		Top 5	Top 5
Supervised baseline	ResNet-50	48.4	80.4
Methods using other label-propagation:			
Pseudo-label	ResNet-50	51.6	82.4
VAT+Entropy Min.	ResNet-50	47.0	83.4
UDA (w. RandAug)	ResNet-50	-	88.5
FixMatch (w. RandAug)	ResNet-50	-	89.1
S4L (Rot+VAT+En. M.)	ResNet-50 (4×)	-	91.2
Methods using representation learning only:			
InstDisc	ResNet-50	39.2	77.4
BigBiGAN	RevNet-50 (4×)	55.2	78.8
PIRL	ResNet-50	57.2	83.8
CPC v2	ResNet-161(∗)	77.9	91.2
SimCLR (ours)	ResNet-50	75.5	87.8
SimCLR (ours)	ResNet-50 (2×)	83.0	91.2
SimCLR (ours)	ResNet-50 (4×)	**85.8**	**92.6**

Table 7. ImageNet accuracy of models trained with few labels.

Chen et al. A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020

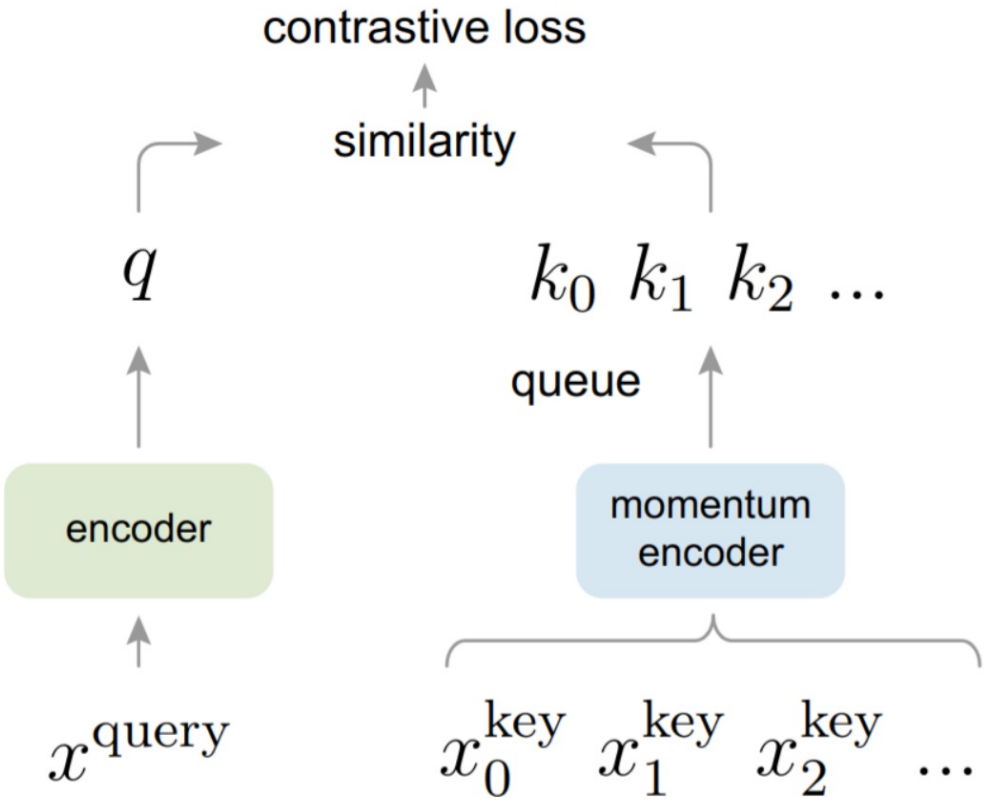
SimCLR



- SimCLR требует обучения с большими батчами
- Это возможно только для распределённого обучения requires training with large
- SimCLR v2 – большие модели, больше слоёв в проекторе

Chen et al. A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020

Momentum Contrast (MoCo)



$$\mathcal{L}_q = -\log \frac{\exp(q \cdot k^+ / \tau)}{\exp(q \cdot k^+ / \tau) + \sum_{k^-} \exp(q \cdot k^- / \tau)}$$

- Цель – уменьшить размер батча для работы на меньшем числе GPU
- Поддерживаем «очередь» из примеров (ключей), которые используем для оценки loss, но не используем их для расчёта градиента
- Для стабилизации «ключи» прогоняем через momentum encoder
- Кодировщик обновляется с большим моментов (m = 0.999)

He et al. Momentum contrast for unsupervised visual representation learning. CVPR 2020

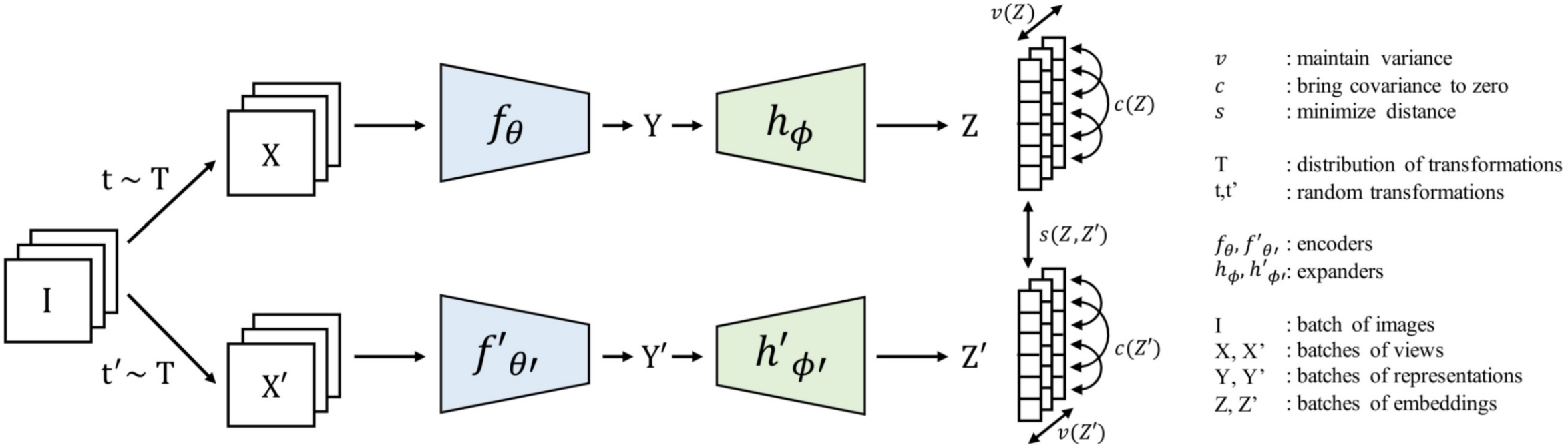
MoCo v2

case	unsup. pre-train MLP	aug+	cos	epochs	batch	ImageNet acc.
MoCo v1 [6]				200	256	60.6
SimCLR [2]	✓	✓	✓	200	256	61.9
SimCLR [2]	✓	✓	✓	200	8192	66.6
MoCo v2	✓	✓	✓	200	256	**67.5**
*results of **longer** unsupervised training follow:*						
SimCLR [2]	✓	✓	✓	1000	4096	69.3
MoCo v2	✓	✓	✓	800	256	**71.1**

- Нелинейный проектор и сильные агументации необходимы для получения хорошего качества
- MoCo обгоняет SimCLR при заметно меньших батчах
- MoCo v2 мы можем обучать на узле с 8×V100 GPUs)

Chen et al. Improved Baselines with Momentum Contrastive Learning. arXiv:2003.04297

VICReg



$$v(Z) = \frac{1}{d}\sum_{i=1}^{d} \max\left(0, 1 - \sqrt{\mathrm{Var}(z^i) + \varepsilon}\right)$$

$$c(Z) = \frac{1}{d}\sum_{i \neq j}[C(Z)]_{ij}^2, \quad C(Z) = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})(z_i - \bar{z})^T$$

$$s(Z, Z') = MSE(Z, Z')$$

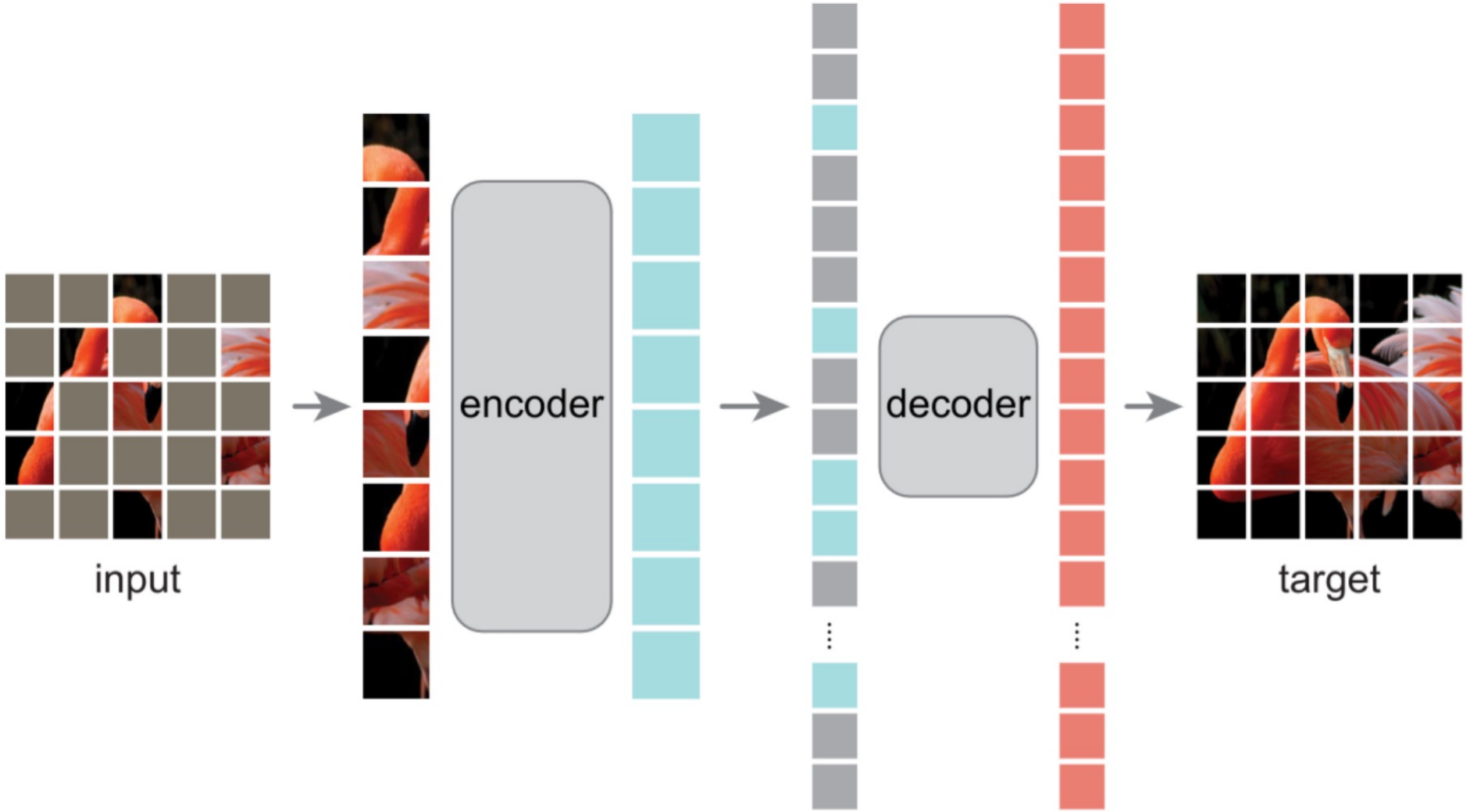
Bardes et al. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. ICLR 2022

VICReg

Method	Linear		Semi-supervised			
	Top-1	Top-5	Top-1		Top-5	
			1%	10%	1%	10%
Supervised	76.5	-	25.4	56.4	48.4	80.4
MoCo He et al. (2020)	60.6	-	-	-	-	-
PIRL Misra & Maaten (2020)	63.6	-	-	-	57.2	83.8
CPC v2 Hénaff et al. (2019)	63.8	-	-	-	-	-
CMC Tian et al. (2019)	66.2	-	-	-	-	-
SimCLR Chen et al. (2020a)	69.3	89.0	48.3	65.6	75.5	87.8
MoCo v2 Chen et al. (2020c)	71.1	-	-	-	-	-
SimSiam Chen & He (2020)	71.3	-	-	-	-	-
SwAV Caron et al. (2020)	71.8	-	-	-	-	-
InfoMin Aug Tian et al. (2020)	73.0	91.1	-	-	-	-
OBoW Gidaris et al. (2021)	73.8	-	-	-	82.9	90.7
BYOL Grill et al. (2020)	74.3	91.6	53.2	68.8	78.4	89.0
SwAV (w/ multi-crop) Caron et al. (2020)	75.3	-	53.9	70.2	78.5	89.9
Barlow Twins Zbontar et al. (2021)	73.2	91.0	55.0	69.7	79.2	89.3
VICReg (ours)	73.2	91.1	54.8	69.5	79.4	89.5

Обучение выполняется на 32×V100 GPUs

Masked Autoencoders



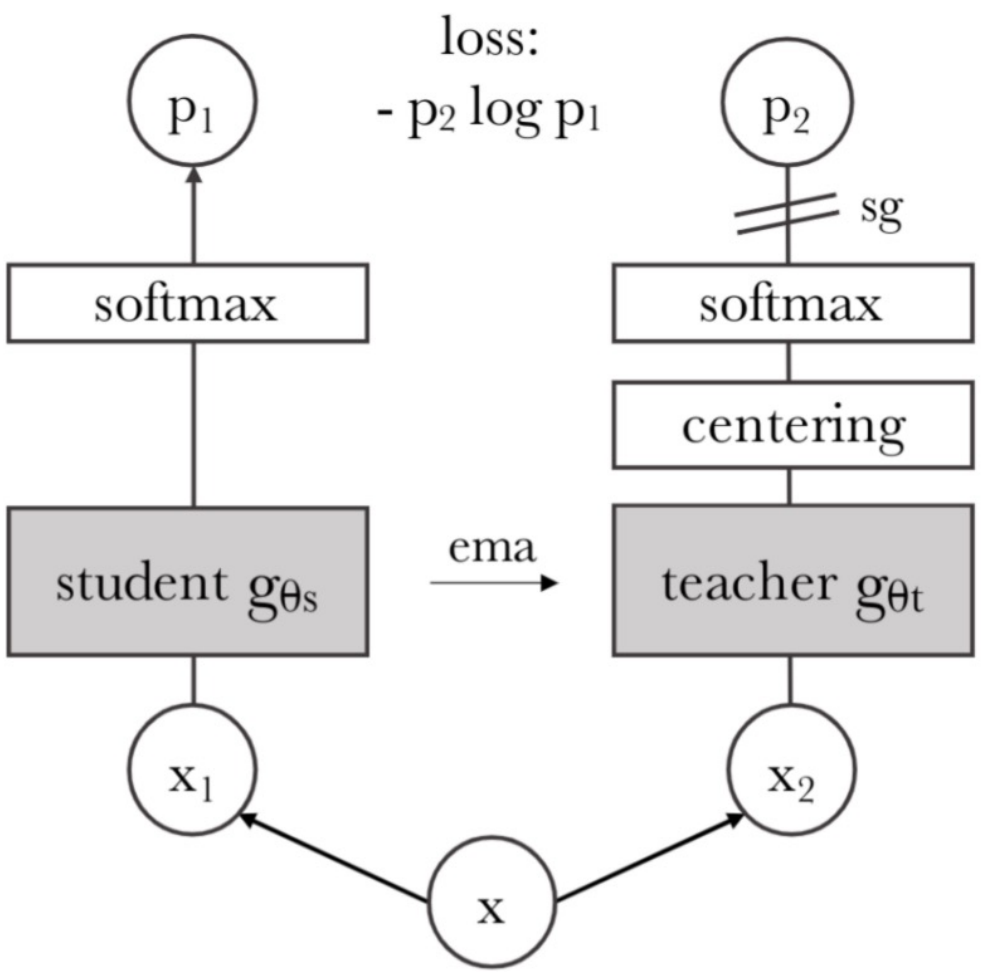
He et al. Masked autoencoders are scalable vision learners. CVPR 2022

Masked Autoencoders

method	pre-train data	ViT-B	ViT-L	ViT-H	ViT-H_{448}
scratch, our impl.	-	82.3	82.6	83.1	-
DINO [5]	IN1K	82.8	-	-	-
MoCo v3 [9]	IN1K	83.2	84.1	-	-
BEiT [2]	IN1K+DALLE	83.2	85.2	-	-
MAE	IN1K	83.6	85.9	86.9	**87.8**

He et al. Masked autoencoders are scalable vision learners. CVPR 2022

DINO



Supervised

DINO

	Random	Supervised	DINO
ViT-S/16	22.0	27.3	45.9
ViT-S/8	21.8	23.7	44.7

Caron et al. Emerging Properties in Self-Supervised Vision Transformers. ICCV 2021

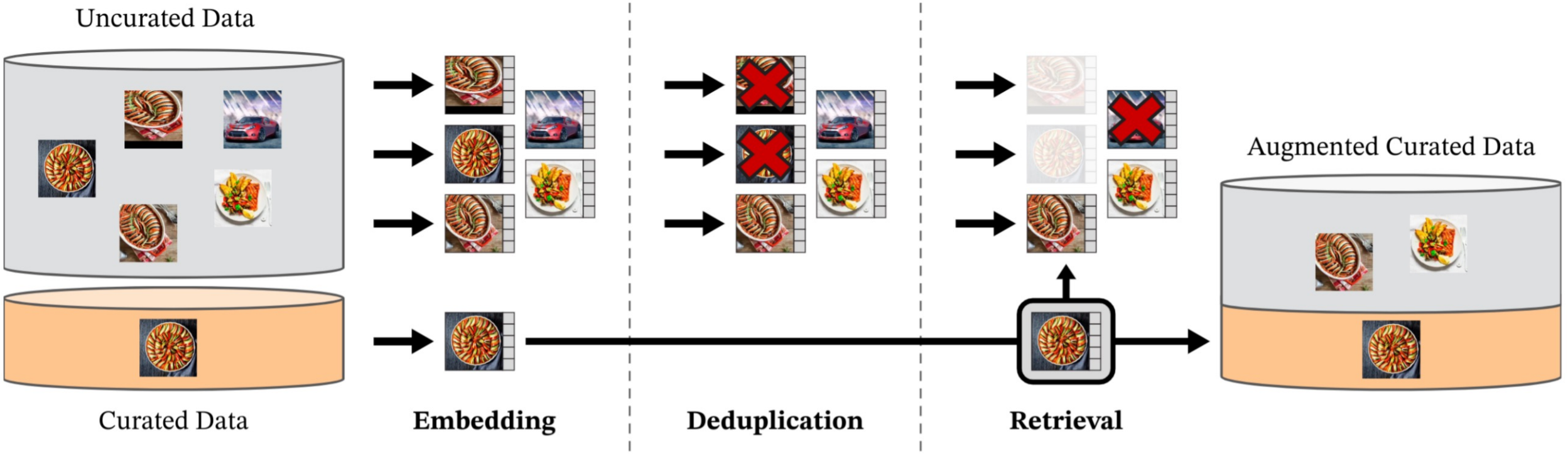
DINO

Method	Arch.	Param.	im/s	Linear	k-NN
Supervised	RN50	23	1237	79.3	79.3
SCLR [12]	RN50	23	1237	69.1	60.7
MoCov2 [15]	RN50	23	1237	71.1	61.9
InfoMin [67]	RN50	23	1237	73.0	65.3
BarlowT [81]	RN50	23	1237	73.2	66.0
OBoW [27]	RN50	23	1237	73.8	61.9
BYOL [30]	RN50	23	1237	74.4	64.8
DCv2 [10]	RN50	23	1237	75.2	67.1
SwAV [10]	RN50	23	1237	**75.3**	65.7
DINO	RN50	23	1237	**75.3**	**67.5**
Supervised	ViT-S	21	1007	79.8	79.8
BYOL* [30]	ViT-S	21	1007	71.4	66.6
MoCov2* [15]	ViT-S	21	1007	72.7	64.4
SwAV* [10]	ViT-S	21	1007	73.5	66.3
DINO	ViT-S	21	1007	**77.0**	**74.5**
Comparison across architectures					
SCLR [12]	RN50w4	375	117	76.8	69.3
SwAV [10]	RN50w2	93	384	77.3	67.3
BYOL [30]	RN50w2	93	384	77.4	–
DINO	ViT-B/16	85	312	78.2	76.1
SwAV [10]	RN50w5	586	76	78.5	67.1
BYOL [30]	RN50w4	375	117	78.6	–
BYOL [30]	RN200w2	250	123	79.6	73.9
DINO	ViT-S/8	21	180	79.7	**78.3**
SCLRv2 [13]	RN152w3+SK	794	46	79.8	73.1
DINO	ViT-B/8	85	63	**80.1**	77.4

Обучение на одном узле 8×V100 GPUs

Caron et al. Emerging Properties in Self-Supervised Vision Transformers. ICCV 2021

DINOv2



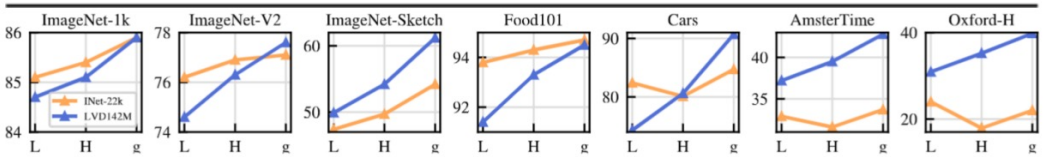
Данные + большая модель (ViT-g, 1.1B params) с дистилляцией + несколько функций потерь и регуляризации + эффективное выполнение. Кодя для обучения и веса выложены в опенсор с коммерческой лицензией

Oquab et al. DINOv2: Learning Robust Visual Features without Supervision. arXiv:2304.07193

DINOv2

Task	Dataset / Split	Images	Retrieval	Retrieved	Final
classification	ImageNet-22k / –	14,197,086	as is	–	14,197,086
classification	ImageNet-22k / –	14,197,086	sample	56,788,344	56,788,344
classification	ImageNet-1k / train	1,281,167	sample	40,997,344	40,997,344
fine-grained classif.	Caltech 101 / train	3,030	cluster	2,630,000	1,000,000
fine-grained classif.	CUB-200-2011 / train	5,994	cluster	1,300,000	1,000,000
fine-grained classif.	DTD / train1	1,880	cluster	1,580,000	1,000,000
fine-grained classif.	FGVC-Aircraft / train	3,334	cluster	1,170,000	1,000,000
fine-grained classif.	Flowers-102 / train	1,020	cluster	1,060,000	1,000,000
fine-grained classif.	Food-101 / train	75,750	cluster	21,670,000	1,000,000
fine-grained classif.	Oxford-IIIT Pet / trainval	3,680	cluster	2,750,000	1,000,000
fine-grained classif.	Stanford Cars / train	8,144	cluster	7,220,000	1,000,000
fine-grained classif.	SUN397 / train1	19,850	cluster	18,950,000	1,000,000
fine-grained classif.	Pascal VOC 2007 / train	2,501	cluster	1,010,000	1,000,000
segmentation	ADE20K / train	20,210	cluster	20,720,000	1,000,000
segmentation	Cityscapes / train	2,975	cluster	1,390,000	1,000,000
segmentation	Pascal VOC 2012 (seg.) / trainaug	1,464	cluster	10,140,000	1,000,000
depth estimation	Mapillary SLS / train	1,434,262	as is	–	1,434,262
depth estimation	KITTI / train (Eigen)	23,158	cluster	3,700,000	1,000,000
depth estimation	NYU Depth V2 / train	24,231	cluster	10,850,000	1,000,000
depth estimation	SUN RGB-D / train	4,829	cluster	4,870,000	1,000,000
retrieval	Google Landmarks v2 / train (clean)	1,580,470	as is	–	1,580,470
retrieval	Google Landmarks v2 / train (clean)	1,580,470	sample	6,321,880	6,321,880
retrieval	AmsterTime / new	1,231	cluster	960,000	960,000
retrieval	AmsterTime / old	1,231	cluster	830,000	830,000
retrieval	Met / train	397,121	cluster	62,860,000	1,000,000
retrieval	Revisiting Oxford / base	4,993	cluster	3,680,000	1,000,000
retrieval	Revisiting Paris / base	6,322	cluster	3,660,000	1,000,000
					142,109,386

Training Data	INet-1k	Im-A	ADE-20k	Oxford-M
INet-22k	85.9	73.5	46.6	62.5
INet-22k \ INet-1k	85.3	70.3	46.2	58.7
Uncurated data	83.3	59.4	48.5	54.3
LVD-142M	85.8	73.9	47.7	64.6



Oquab et al. DINOv2: Learning Robust Visual Features without Supervision. arXiv:2304.07193

DINOv2

Method	Arch.	Data	Text sup.	kNN	linear		
				val	val	ReaL	V2
Weakly supervised							
CLIP	ViT-L/14	WIT-400M	✓	79.8	84.3	88.1	75.3
CLIP	ViT-L/14_{336}	WIT-400M	✓	80.5	85.3	88.8	75.8
SWAG	ViT-H/14	IG3.6B	✓	82.6	85.7	88.7	77.6
OpenCLIP	ViT-H/14	LAION	✓	81.7	84.4	88.4	75.5
OpenCLIP	ViT-G/14	LAION	✓	83.2	86.2	89.4	77.2
EVA-CLIP	ViT-g/14	custom*	✓	**83.5**	86.4	89.3	77.4
Self-supervised							
MAE	ViT-H/14	INet-1k	×	49.4	76.6	83.3	64.8
DINO	ViT-S/8	INet-1k	×	78.6	79.2	85.5	68.2
SEERv2	RG10B	IG2B	×	–	79.8	–	–
MSN	ViT-L/7	INet-1k	×	79.2	80.7	86.0	69.7
EsViT	Swin-B/W=14	INet-1k	×	79.4	81.3	87.0	70.4
Mugs	ViT-L/16	INet-1k	×	80.2	82.1	86.9	70.8
iBOT	ViT-L/16	INet-22k	×	72.9	82.3	87.5	72.4
DINOv2	ViT-S/14	LVD-142M	×	79.0	81.1	86.6	70.9
	ViT-B/14	LVD-142M	×	82.1	84.5	88.3	75.1
	ViT-L/14	LVD-142M	×	**83.5**	86.3	89.5	78.0
	ViT-g/14	LVD-142M	×	**83.5**	**86.5**	**89.6**	**78.4**

План лекции

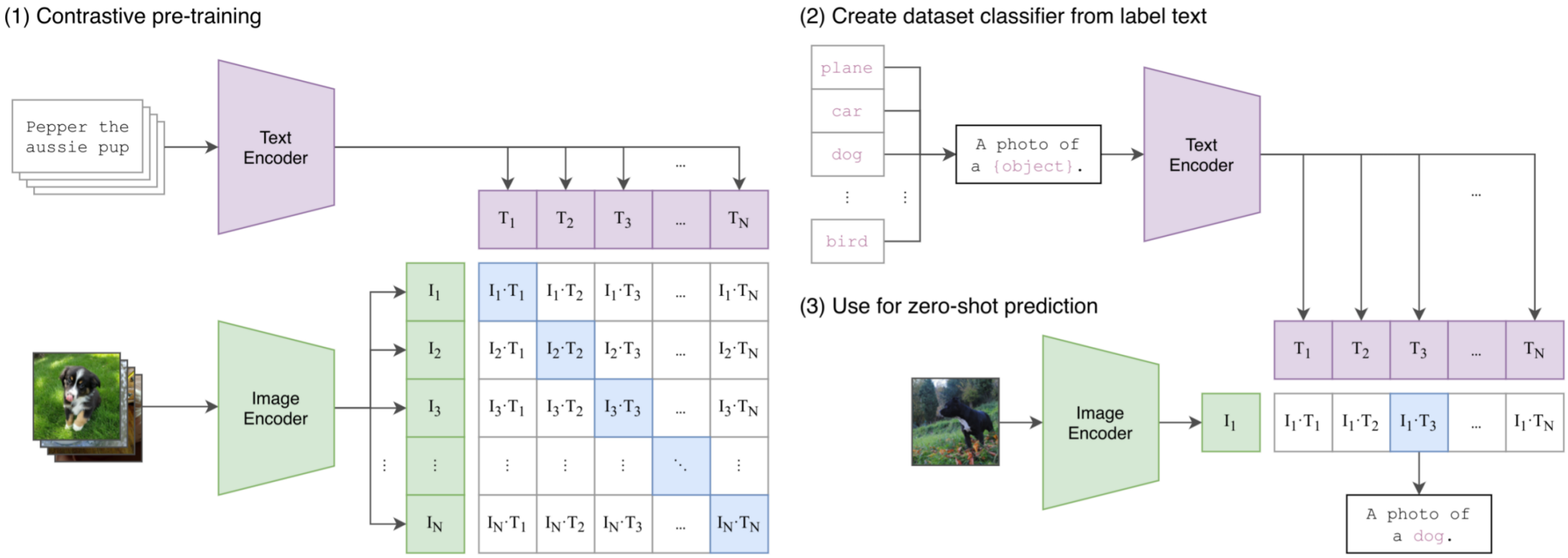
1. Введение
2. Прокси-задачи
3. Констранстное обучение и маскирование
4. Фундаментальные модели

Что такое фундаментальная модель?

Модель называют «фундаментальной», если:

- Работает с несколькими модальностями, например, текстом и изображениями
- Решают несколько задач в разных доменах, например, классификацию, сегментацию, описание, ответы на вопросы
- Работает с запросами, т.е. поддерживает несколько вариантов запросов относительно анализируемой информации
- Хорошо работает без тонкой настройки на целевой бенчмарк

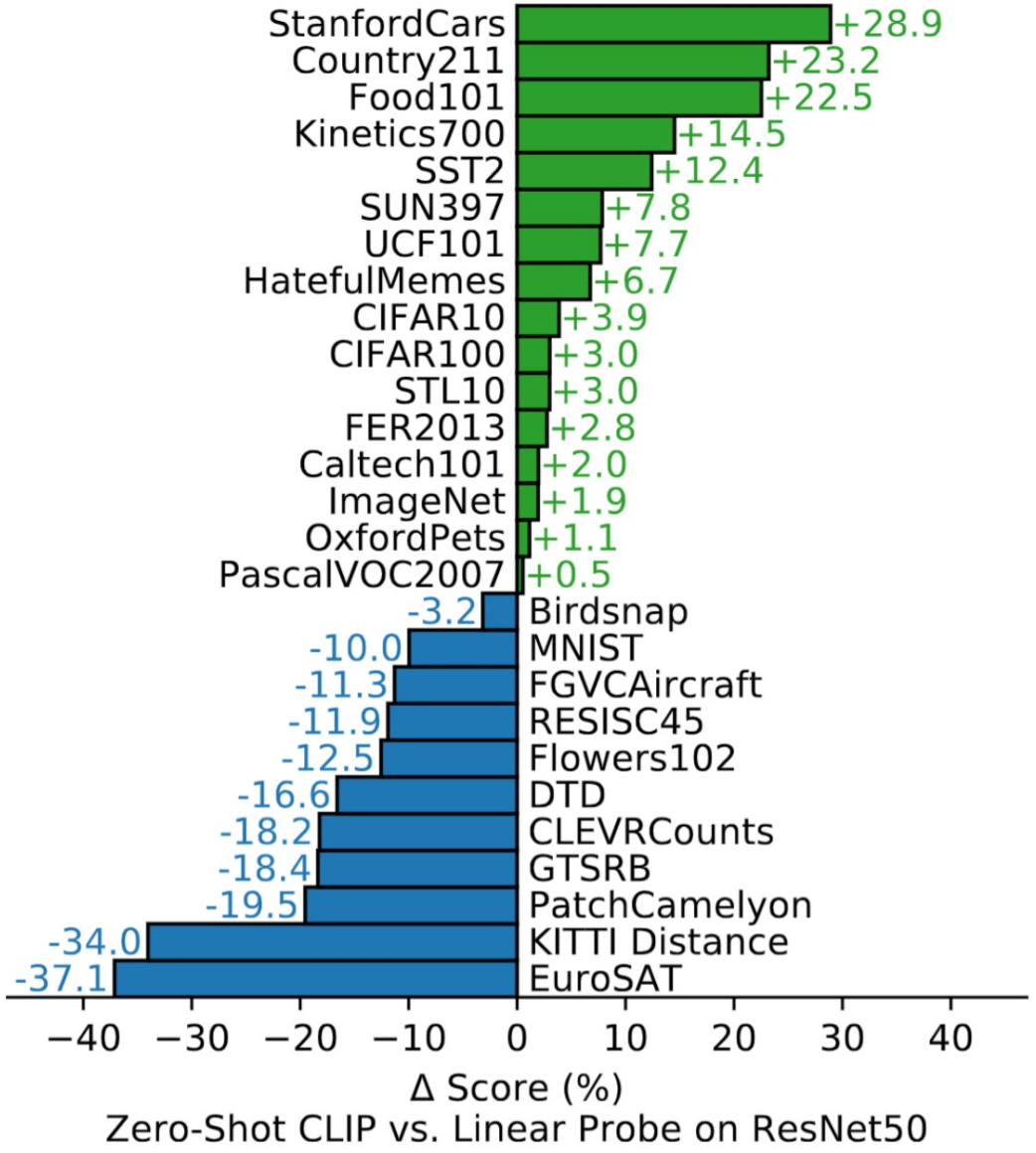
CLIP



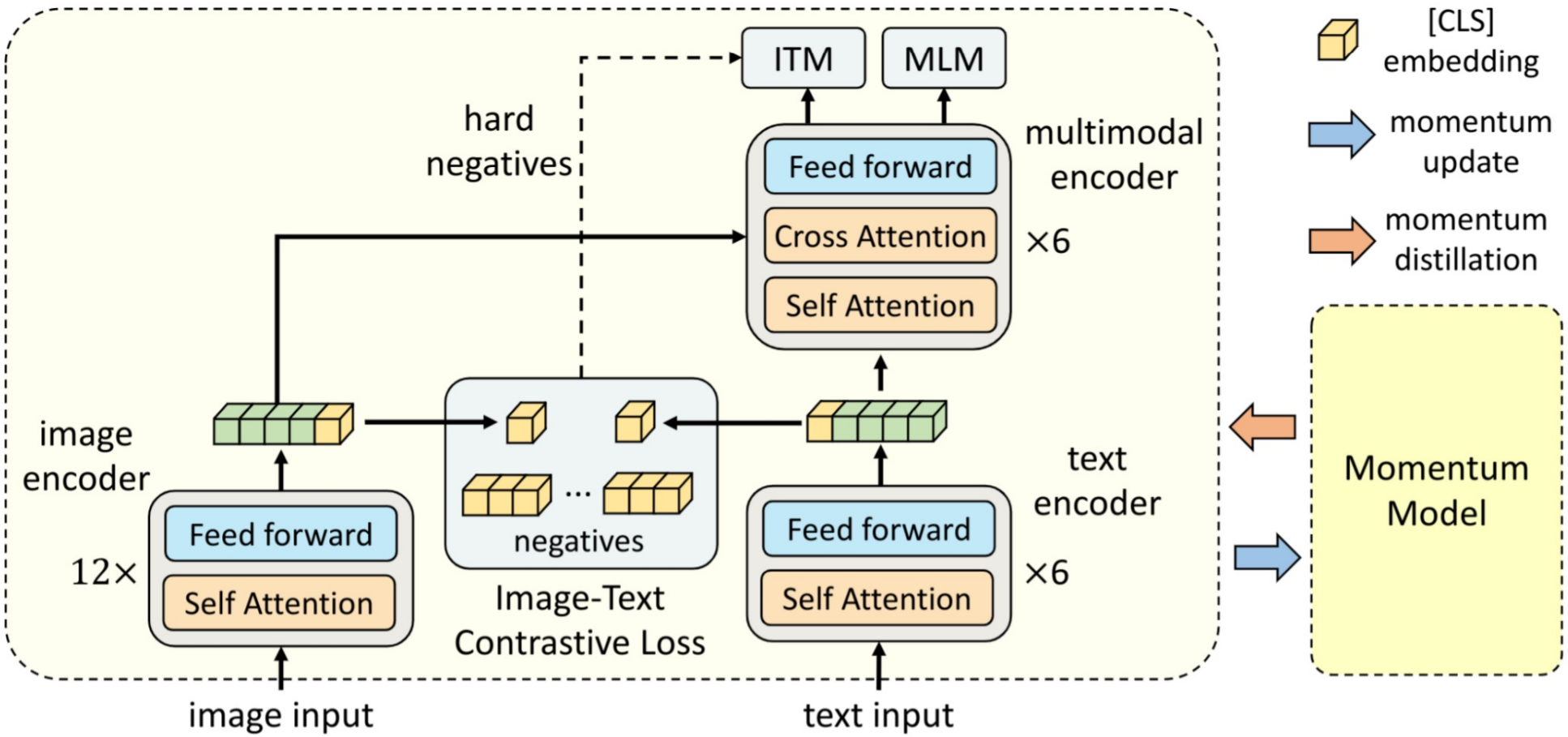
400M (image, text) пар, 500×V100 GPUs для предобучения

Radford et al. Learning transferable visual models from natural language supervision. ICML 2021

CLIP zero-shot results



ALBEF



Loss: contrastive loss + image-text matching + masked language modelling

Li et al. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. NeurIPS 2021

ALBEF pseudo-targets

masked language modelling ↓

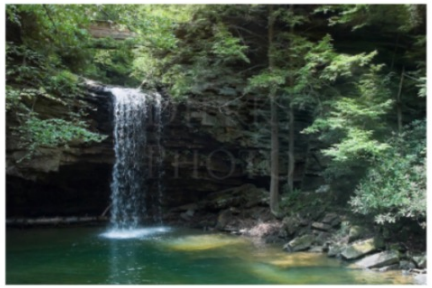
"polar bear in the [MASK]"

GT: wild
Top-5 pseudo-targets:
1. zoo
2. pool
3. water
4. pond
5. wild

"a man [MASK] along a road in front of nature in summer"

GT: standing
Top-5 pseudo-targets:
1. walks
2. walking
3. runs
4. running
5. goes

"a [MASK] waterfall in the deep woods"



GT: remote
Top-5 pseudo-targets:
1. small
2. beautiful
3. little
4. secret
5. secluded

GT: breakdown of the car on the road
Top-5 pseudo-targets:
1. young woman get out of the car near the road
2. a woman inspects her damaged car under a tree
3. a woman looking into a car after locking her keys inside
4. young woman with a broken car calling for help
5. breakdown of the car on the road

GT: the harbor a small village
Top-5 pseudo-targets:
1. the harbour with boats and houses
2. replica of the sailing ship in the harbour
3. ships in the harbor of the town
4. the harbor a small village
5. boats lined up alongside the geographical feature category in the village

image-text matching ↑

Сравнение ALBEF с CLIP

Method	# Pre-train Images	Flickr30K (1K test set) TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10
UNITER [2]	4M	83.6	95.7	97.7	68.7	89.2	93.9
CLIP [6]	400M	88.0	98.7	99.4	68.7	90.6	95.2
ALIGN [7]	1.2B	88.6	98.7	99.7	75.7	93.8	96.8
ALBEF	4M	90.5	98.8	99.7	76.8	93.7	96.7
ALBEF	14M	**94.1**	**99.5**	**99.7**	**82.8**	**96.3**	**98.1**

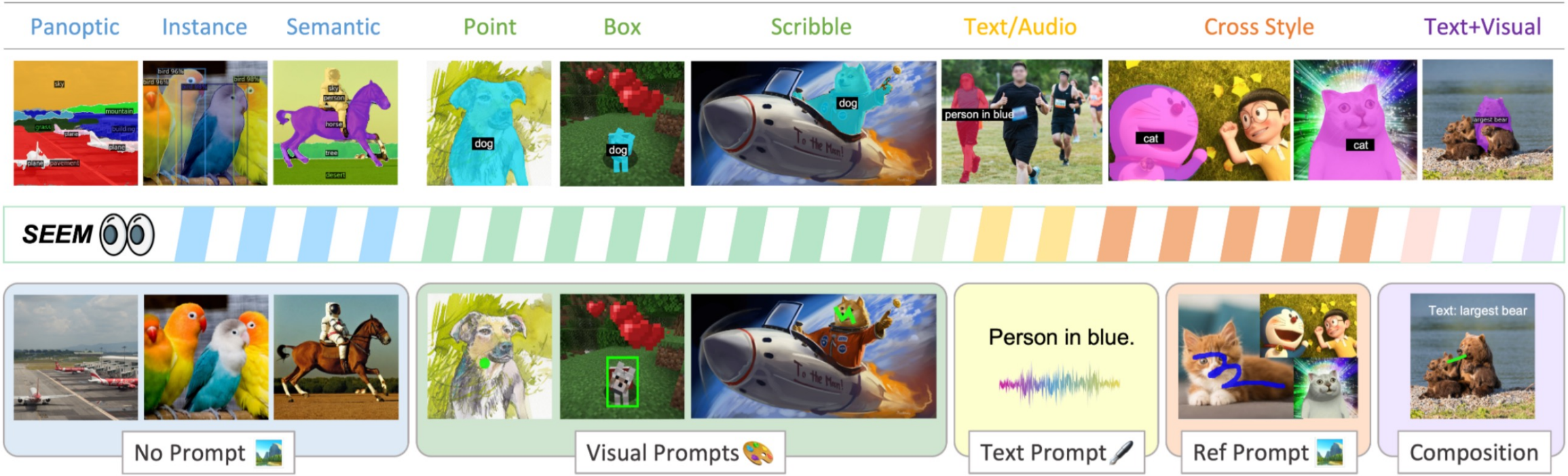
Table 3: Zero-shot image-text retrieval results on Flickr30K.

Method	VQA		NLVR^2		SNLI-VE	
	test-dev	test-std	dev	test-P	val	test
VisualBERT [13]	70.80	71.00	67.40	67.00	-	-
VL-BERT [10]	71.16	-	-	-	-	-
LXMERT [1]	72.42	72.54	74.90	74.50	-	-
12-in-1 [12]	73.15	-	-	78.87	-	76.95
UNITER [2]	72.70	72.91	77.18	77.85	78.59	78.28
VL-BART/T5 [54]	-	71.3	-	73.6	-	-
ViLT [21]	70.94	-	75.24	76.21	-	-
OSCAR [3]	73.16	73.44	78.07	78.36	-	-
VILLA [8]	73.59	73.67	78.39	79.30	79.47	79.03
ALBEF (4M)	74.54	74.70	80.24	80.50	80.14	80.30
ALBEF (14M)	**75.84**	**76.04**	**82.55**	**83.14**	**80.80**	**80.91**

Table 4: Comparison with state-of-the-art methods on downstream vision-language tasks.

8×A100 GPUs для предобучения

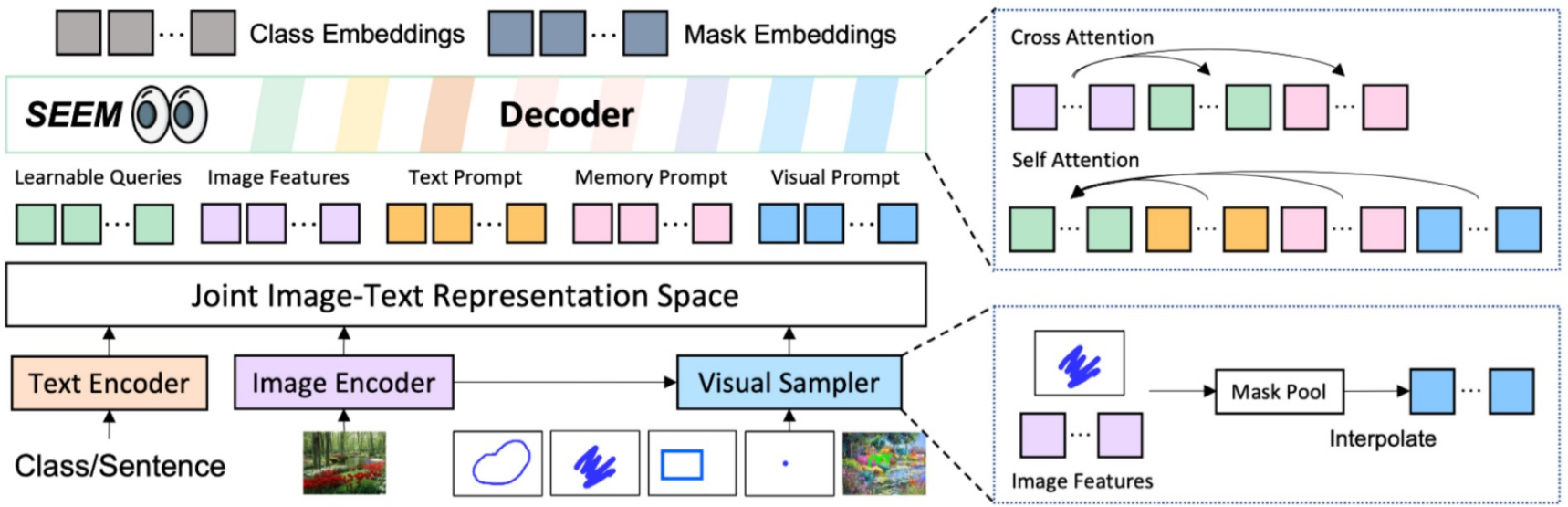
SEEM



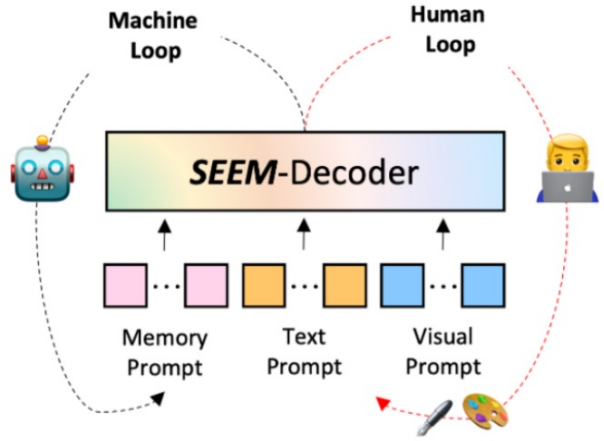
Panoptic + referring + interactive segmentation

Zou et al. Segment Everything Everywhere All at Once. NeurIPS 2023

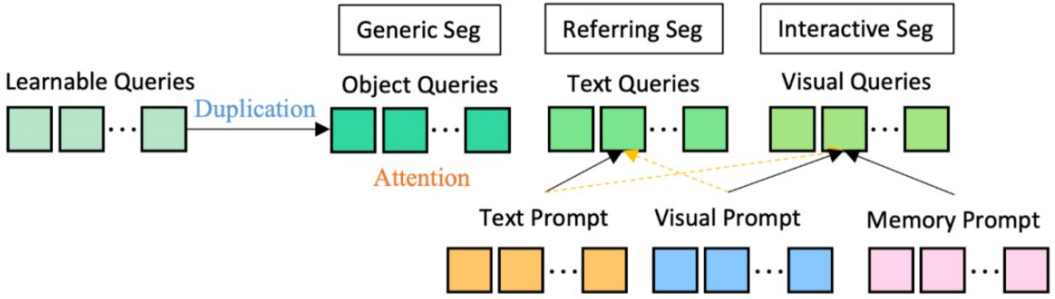
SEEM



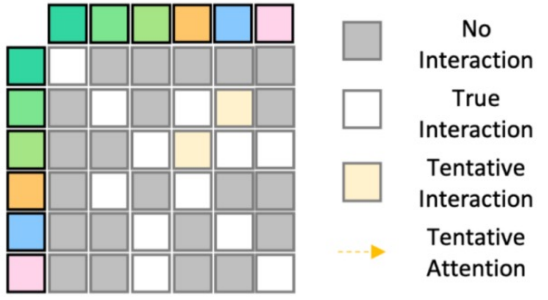
(a) Model Architecture



(b) Human-Model Interaction



(a) Queries and Prompt Interaction



(b) Self-Attention Mask

Zou et al. Segment Everything Everywhere All at Once. NeurIPS 2023

SEEM

Method	Segmentation Data	Type	Generic Segmentation COCO			Referring Segmentation RefCOCOg			Interactive Segmentation PascalVOC					
			PQ	mAP	mIoU	cIoU	mIoU	AP50	5-NoC85	10-NoC85	20-NoC85	5-NoC90	10-NoC90	20-NoC90
Mask2Former (T) [6]	COCO (0.12M)	Segmentation	53.2	43.3	63.2	-	-	-	-	-	-	-	-	-
Mask2Former (B) [6]	COCO (0.12M)		56.4	46.3	67.1	-	-	-	-	-	-	-	-	-
Mask2Former (L) [6]	COCO (0.12M)		57.8	48.6	67.4	-	-	-	-	-	-	-	-	-
Pano/SegFormer (B) [45]	COCO (0.12M)		55.4	*	*	-	-	-	-	-	-	-	-	-
LAVT (B) [53]	Ref-COCO (0.03M)		-	-	-	61.2	*	*	-	-	-	-	-	-
PolyFormer (B) [17]	Ref-COCO+VG+... (0.16M)		-	-	-	69.3	*	*	-	-	-	-	-	-
PolyFormer (L) [17]	Ref-COCO+VG+... (0.16M)		-	-	-	71.1	*	*	-	-	-	-	-	-
RITM (<T) [18]	COCO+LVIS (0.12M)	Interactive	-	-	-	-	-	-	*	*	2.19	*	*	2.57
PseudoClick (<T) [54]	COCO (0.12M)		-	-	-	-	-	-	*	*	1.94	*	*	2.25
FocalClick (T) [21]	COCO (0.12M)		-	-	-	-	-	-	*	*	2.97	*	*	3.52
FocalClick (B) [21]	COCO (0.12M)		-	-	-	-	-	-	*	*	2.46	*	*	2.88
SimpleClick (B) [20]	COCO+LVIS (0.12M)		-	-	-	-	-	-	1.75	1.93	2.06	1.94	2.19	2.38
SimpleClick (L) [20]	COCO+LVIS (0.12M)		-	-	-	-	-	-	1.52	1.64	1.72	1.67	1.84	1.96
SimpleClick (H) [20]	COCO+LVIS (0.12M)		-	-	-	-	-	-	1.51	1.64	1.76	1.64	1.83	1.98
UViM (L) [55]	COCO (0.12M)	Generalist	45.8	*	*	-	-	-	-	-	-	-	-	-
Pix2Seq v2 (B) [56]	COCO (0.12M)		-	38.2	-	-	-	-	-	-	-	-	-	-
X-Decoder (T) [11]	COCO (0.12M)		52.6	41.3	62.4	59.8	*	*	-	-	-	-	-	-
X-Decoder (B) [11]	COCO (0.12M)		56.2	45.8	66.0	64.5	*	*	-	-	-	-	-	-
X-Decoder (L) [11]	COCO (0.12M)		56.9	46.7	67.5	64.6	*	*	-	-	-	-	-	-
UNINEXT (T) [48]	Image+Video (3M)		-	44.9	-	70.0	*	*	-	-	-	-	-	-
UNINEXT (L) [48]	Image+Video (3M)		-	49.6	-	73.4	*	*	-	-	-	-	-	-
Painter (L) [57]	COCO+ADE+NYUv2 (0.16M)		43.4	*	*	-	-	-	-	-	-	-	-	-
#SegGPT (L) [50]	COCO+ADE+NYUv2 (0.16M)		34.4	*	*	-	-	-	-	-	-	-	-	-
#SAM (B) [36]	SAM (11M)		-	-	-	-	-	-	2.47	2.65	3.28	2.23	3.13	4.12
#SAM (L) [36]	SAM (11M)		-	-	-	-	-	-	1.85	2.15	2.60	2.01	2.46	3.12
#SAM (H) [36]	SAM (11M)		-	-	-	-	-	-	1.82	2.13	2.55	1.98	2.43	3.11
SEEM (T)	COCO+LVIS (0.12M)		50.8	39.7	62.2	60.9	65.7	74.8	1.72	2.30	3.37	1.97	2.83	4.41
SEEM (B)	COCO+LVIS (0.12M)		56.1	46.4	66.3	65.0	69.6	78.2	1.56	2.04	2.93	1.77	2.47	3.79
SEEM (L)	COCO+LVIS (0.12M)		57.5	47.7	67.6	65.6	70.3	78.9	1.51	1.95	2.77	1.71	2.36	3.61
SEEM (T)	COCO+LVIS (0.12M)	Composition	-	-	-	70.4	71.7	82.1	1.72	2.28	3.32	1.97	2.82	4.37
SEEM (B)	COCO+LVIS (0.12M)		-	-	-	76.2	77.8	87.8	1.56	2.03	2.91	1.77	2.46	3.76
SEEM (L)	COCO+LVIS (0.12M)		-	-	-	75.1	76.9	86.8	1.52	1.97	2.81	1.72	2.38	3.64

Zou et al. Segment Everything Everywhere All at Once. NeurIPS 2023

Florence-2

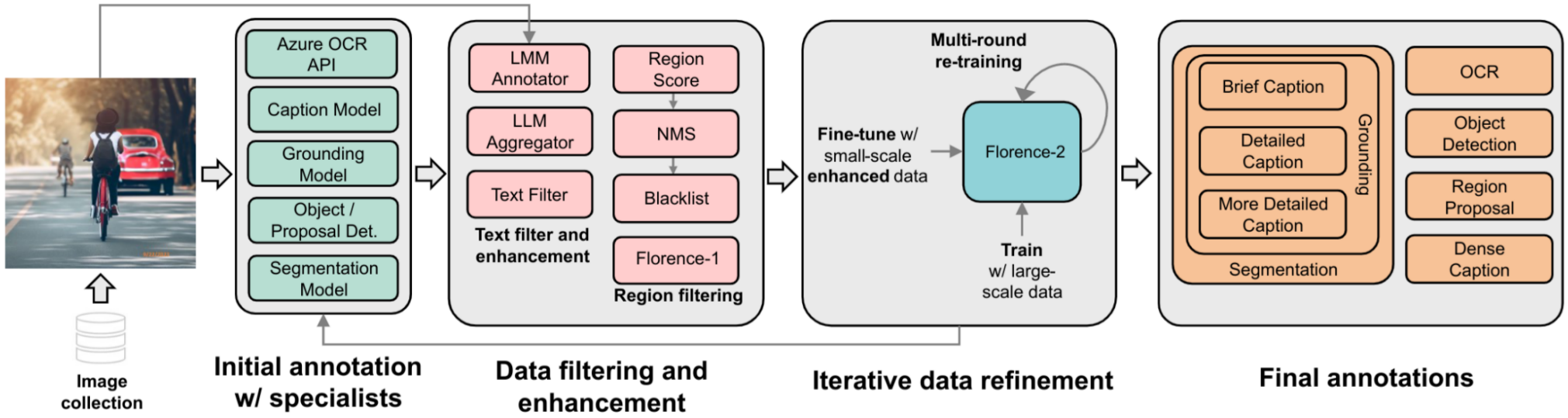
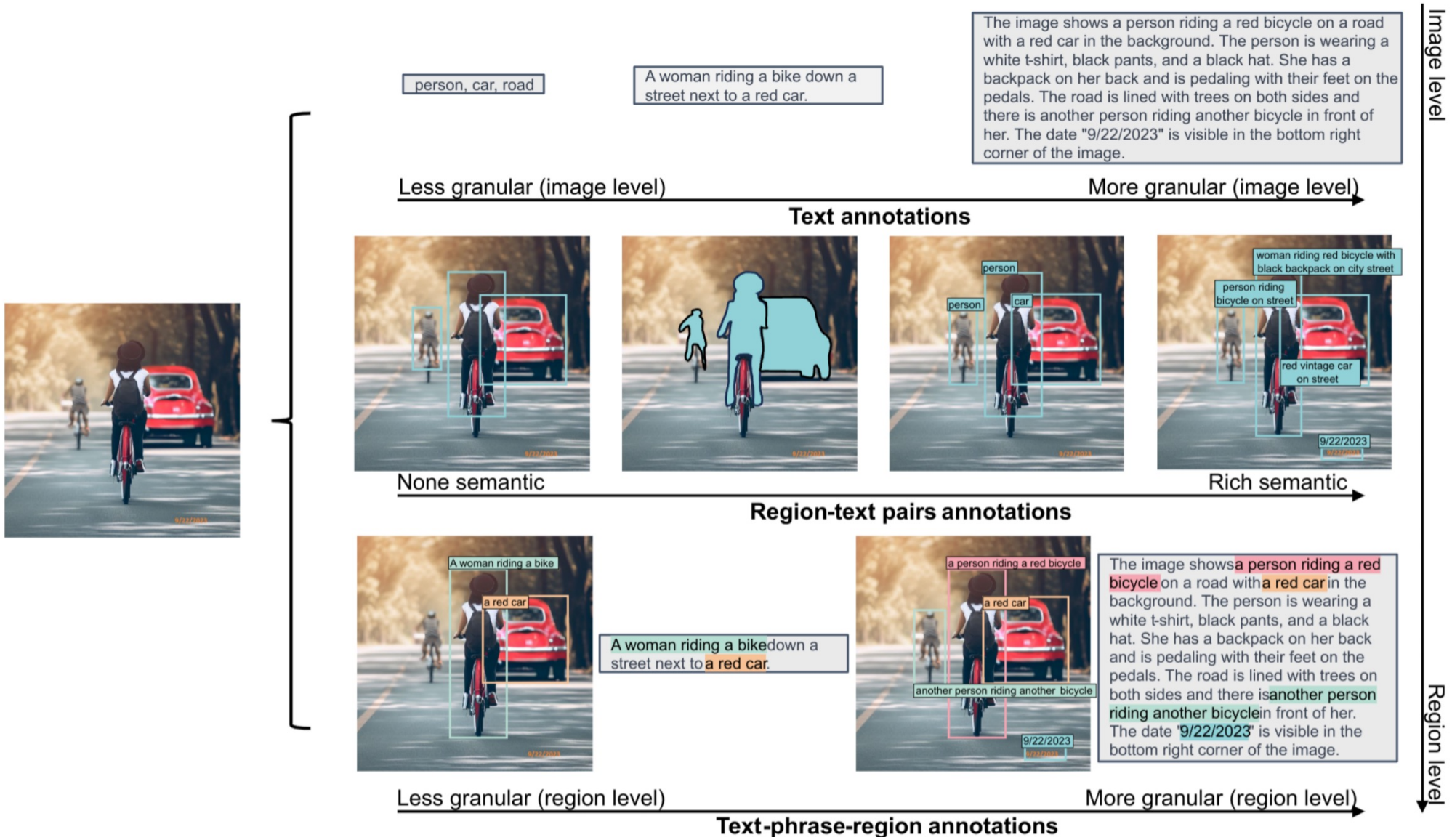


Figure 3. ***Florence-2* data engine** consists of three essential phrases: (1) initial annotation employing specialist models, (2) data filtering to correct errors and remove irrelevant annotations, and (3) an iterative process for data refinement. Our final dataset (***FLD-5B***) of over **5B** annotations contains **126M** images, **500M** text annotations, **1.3B** region-text annotations, and **3.6B** text-phrase-region annotations.

Dataset	Rep. Model	#Images	#Annotations	Spatial hierarchy	Semantics granularity
JFT300M [21]	ViT	300M	300M	Image-level	Coarse
WIT [64]	CLIP	400M	400M	Image-level	Coarse
SA-1B [32]	SAM	11M	1B	Region-level	Non-semantic
GrIT [60]	Kosmos-2	91M	137M	Image & Region-level	Fine-grained
M3W [2]	Flamingo	185M	43.3M*	Multi-image-level	Fine-grained
FLD-5B (ours)	*Florence-2* (ours)	126M	5B	Image & Region-level	Coarse to fine-grained

Xiao et al. Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks. arXiv:2311.06242

Florence 2


person, car, road
A woman riding a bike down a street next to a red car.
The image shows a person riding a red bicycle on a road with a red car in the background. The person is wearing a white t-shirt, black pants, and a black hat. She has a backpack on her back and is pedaling with their feet on the pedals. The road is lined with trees on both sides and there is another person riding another bicycle in front of her. The date "9/22/2023" is visible in the bottom right corner of the image.
Less granular (image level)
More granular (image level)
Text annotations
person
car
person
woman riding red bicycle with black backpack on city street
person riding bicycle on street
red vintage car on street
9/22/2023
None semantic
Rich semantic
Region-text pairs annotations
A woman riding a bike
a red car
A woman riding a bikedown a street next toa red car.
a person riding a red bicycle
a red car
another person riding another bicycle
9/22/2023
The image showsa person riding a red bicycle on a road witha red car in the background. The person is wearing a white t-shirt, black pants, and a black hat. She has a backpack on her back and is pedaling with their feet on the pedals. The road is lined with trees on both sides and there isanother person riding another bicyclein front of her. The date '9/22/2023' is visible in the bottom right corner of the image.
Less granular (region level)
More granular (region level)
Text-phrase-region annotations
Image level
Region level

Florence 2

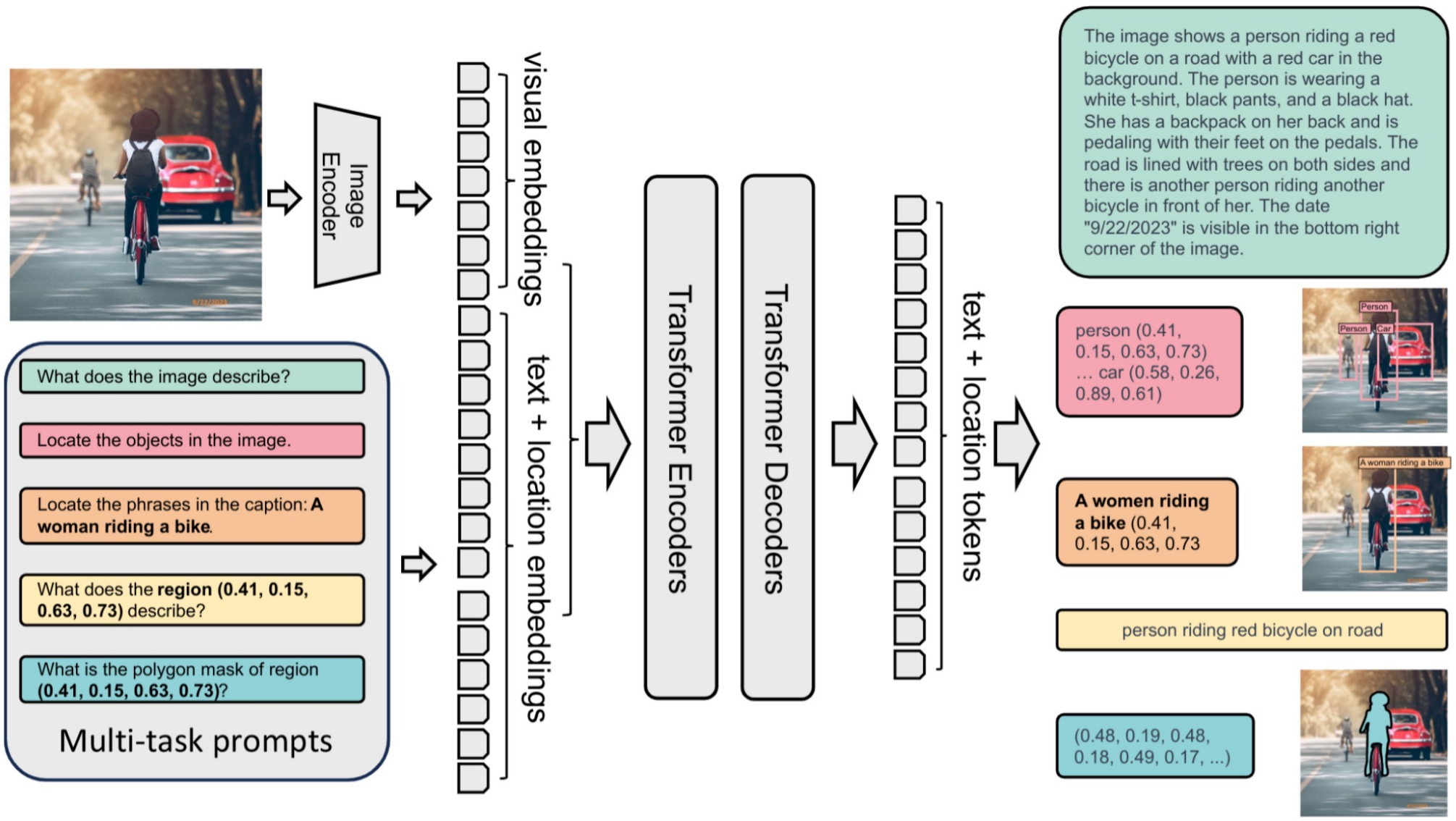
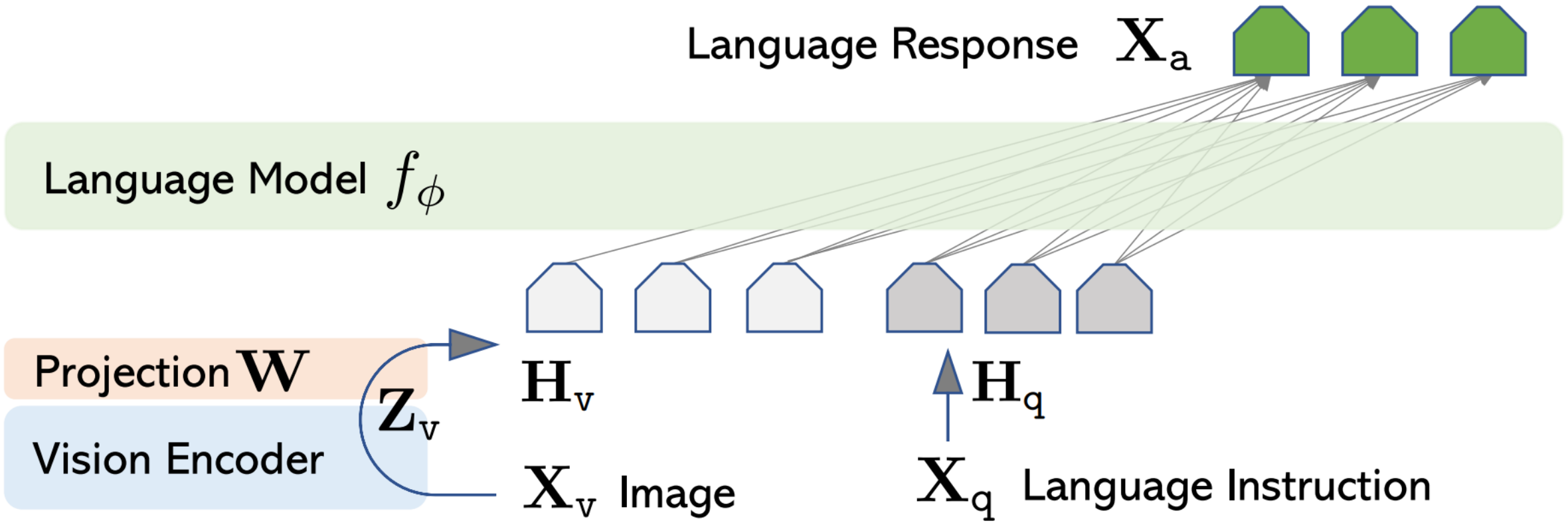


Image Encoder
visual embeddings
text + location embeddings
Transformer Encoders
Transformer Decoders
text + location tokens
What does the image describe?
Locate the objects in the image.
Locate the phrases in the caption: A woman riding a bike.
What does the region (0.41, 0.15, 0.63, 0.73) describe?
What is the polygon mask of region (0.41, 0.15, 0.63, 0.73)?
Multi-task prompts
The image shows a person riding a red bicycle on a road with a red car in the background. The person is wearing a white t-shirt, black pants, and a black hat. She has a backpack on her back and is pedaling with their feet on the pedals. The road is lined with trees on both sides and there is another person riding another bicycle in front of her. The date "9/22/2023" is visible in the bottom right corner of the image.
person (0.41, 0.15, 0.63, 0.73) … car (0.58, 0.26, 0.89, 0.61)
Person
Person
Car
A women riding a bike (0.41, 0.15, 0.63, 0.73
A woman riding a bike
person riding red bicycle on road
(0.48, 0.19, 0.48, 0.18, 0.49, 0.17, ...)

LLaVA: Large Language and Vision Assistant



https://github.com/haotian-liu/LLaVA

Резюме

- Использование неразмеченных данных выгодно, т.к. Их много а размечать долго
- Первый подход к само-обучению заключался в эвристических прокси-задачах
- Контрастное обучение и маскирование – мощные подходы к само-обучению
- Мы рассмотрели определение и несколько примеров фундаментальных моделей