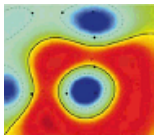
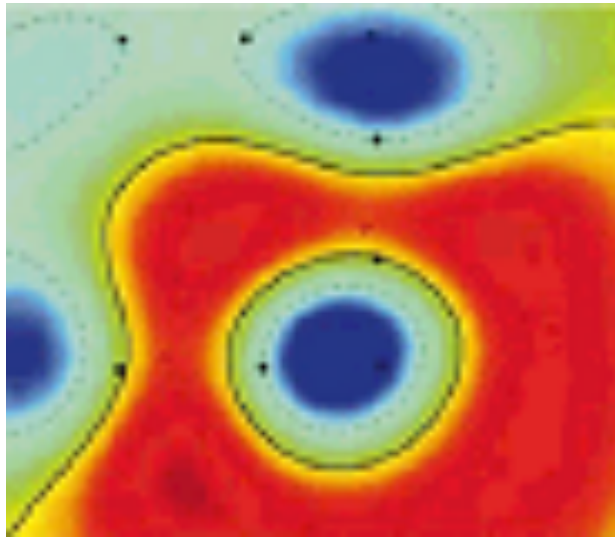


SoSe 2023

Deep Learning 2

Technische Universität Berlin



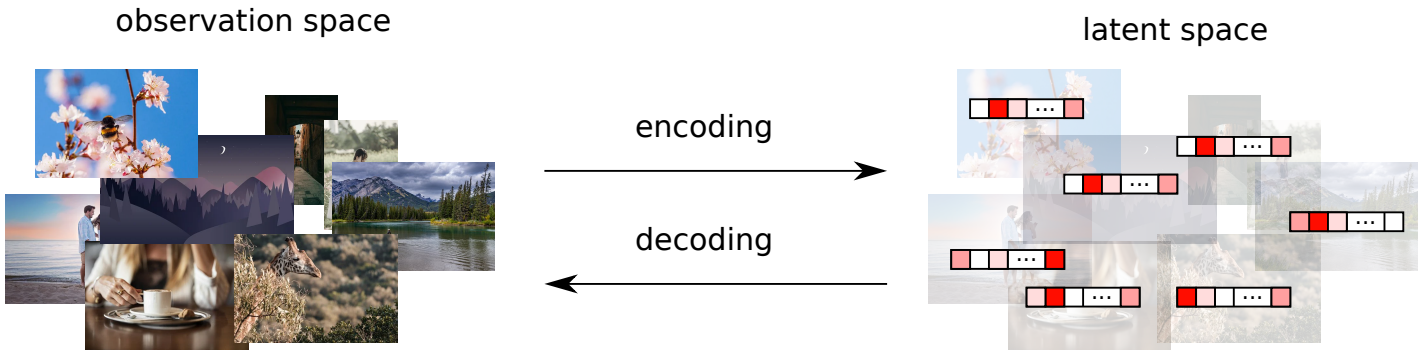
Lecture 1

Representation Learning

Motivation

- Supervised learning has several disadvantages.
  - Labeling data is costly.
  - In some domains, labeling needs annotations from experts (e.g histopathology).
- Often, we have a lot of unlabeled data available (e.g. images/text on the internet).
- Can we make use of patterns in this data?

Representation Learning



- **Idea:** Learn semantic representations of data from patterns in observation space.
  - By learning from large-scale amount of data, we aim to learn a better/more general-purpose feature extractor than from limited supervised data.
  - Through clustering the latent space, we can discover groups or independent components (disentangling the data).
  - By mapping the data into a vector space, we can correlate/fuse data to another data modality (e.g. text).
  - Through comparing similarity in the learned vector space, we can perform semantic search.

Pretext Tasks

Jigsaw Puzzles [Noroozi, 2016]

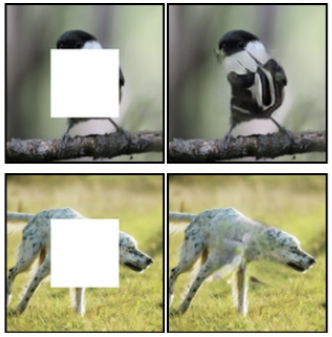
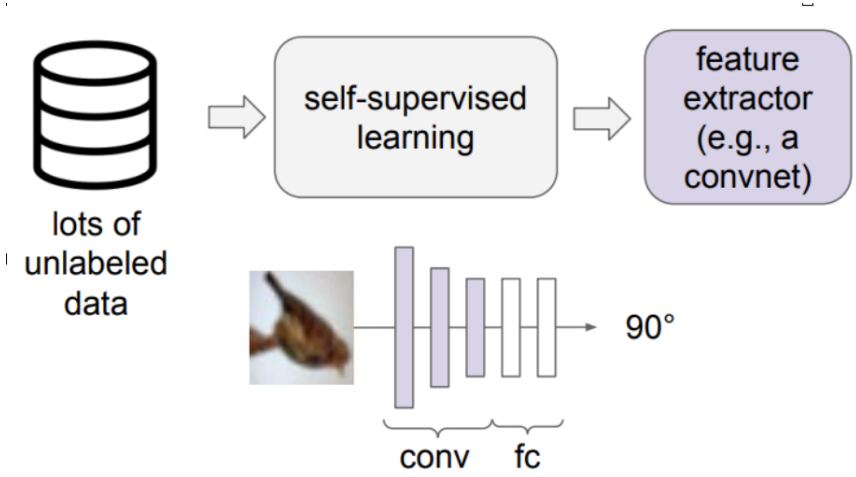


Image Inpainting [Pathak et al., 2016]

Colorization [Zhang et al. 2016]

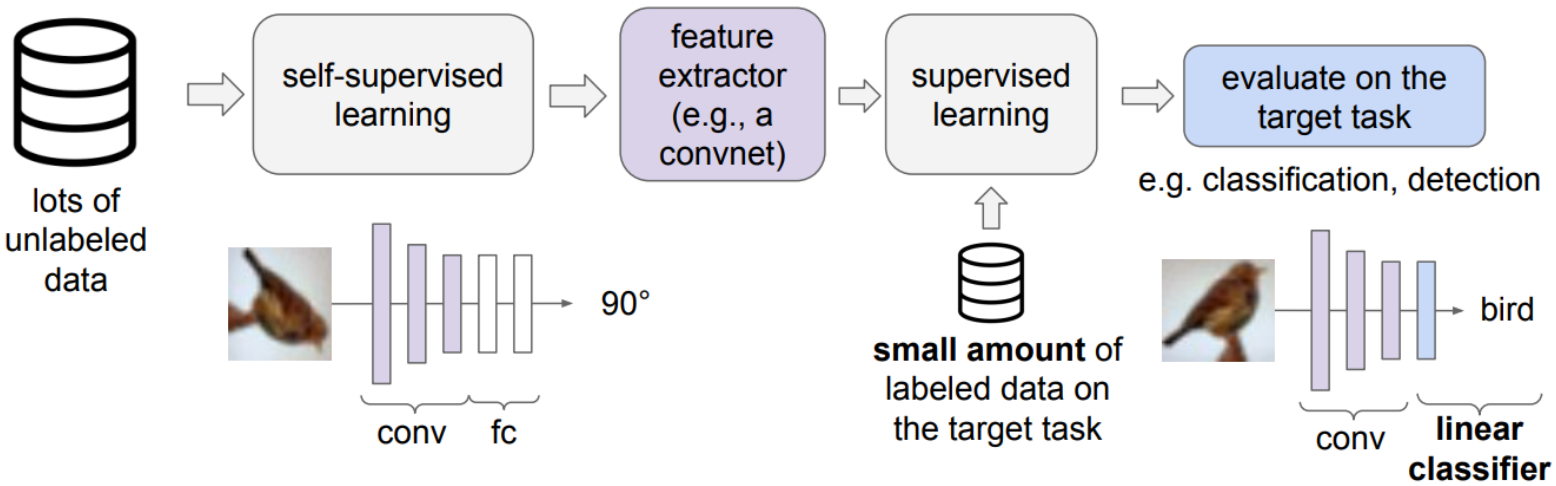
- We need a learning objective that does not require labels.
- Such an auxiliary task is called **Pretext task**.
- The pretext task is usually performed on a property that is inherent in the dataset itself.
- Examples:
  - Predict permutation of image patches (Jigsaw Puzzles) [8].
  - Predict missing contents of an image (Image Inpainting) [9].
  - Colorize a grayscale image [11].
  - Predict the rotation of an image [4].

Self-Supervised Learning



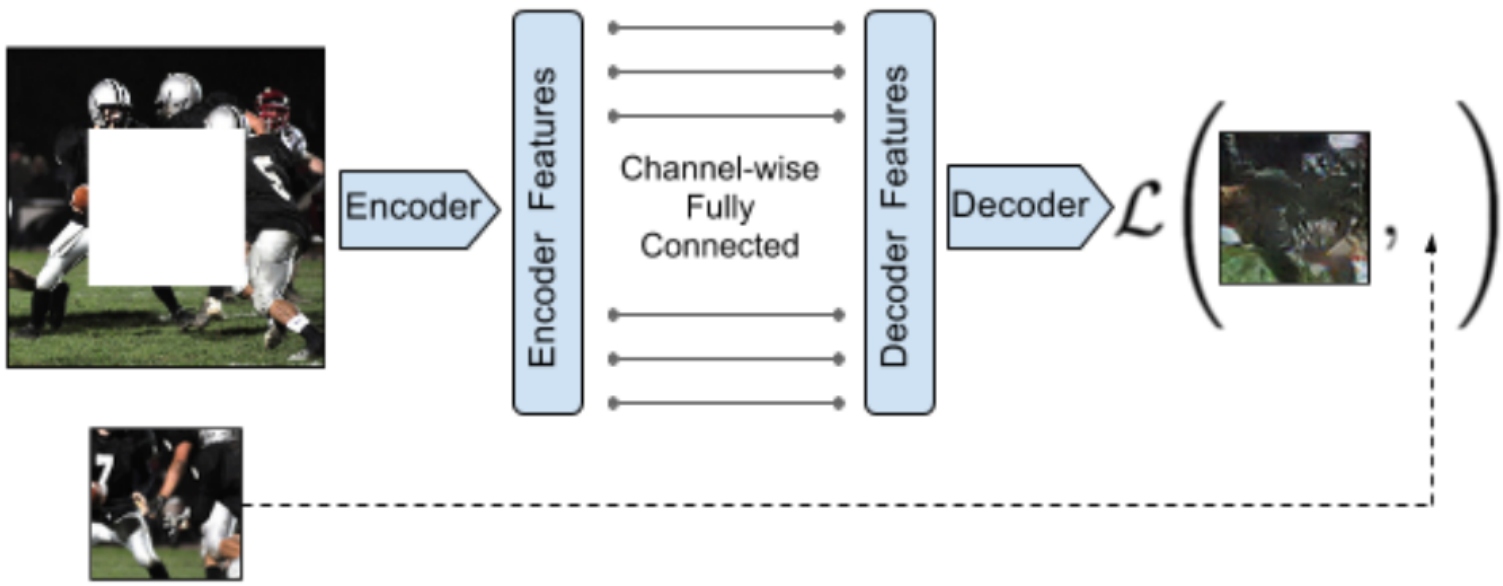
- Learning a good feature extractor (e.g. convolutional network) from unlabeled data before finetuning on a downstream task is called **self-supervised learning**.
- The feature extractor/encoder maps our observation space (e.g. images) to the latent space.
- Self-supervised learning separates feature learning from learning the downstream task while supervised learning usually learns both together.

Transfer to Downstream Tasks



- After pretraining, we freeze the parameters of the feature extractor and only finetune a linear head on the supervised downstream task. This is commonly called **linear probing** or **linear evaluation**.
- By having a general-purpose feature extractor, the computation for training the model on the downstream task gets drastically reduced.

Inpainting



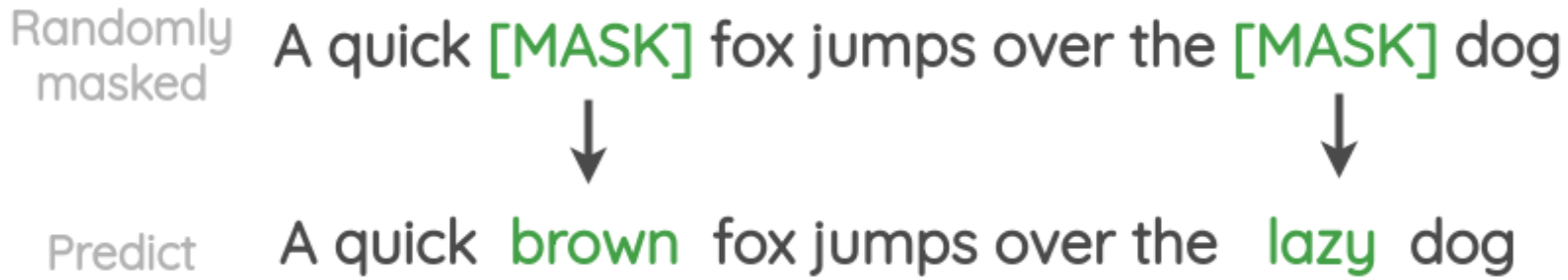
- Idea: occlude a random part of the image with a mask [9].
- The model has to predict the missing contents behind the mask.
- The model is optimized with a mean squared error and an adversarial loss (following lecture).

Inpainting

Input Context Context Encoder

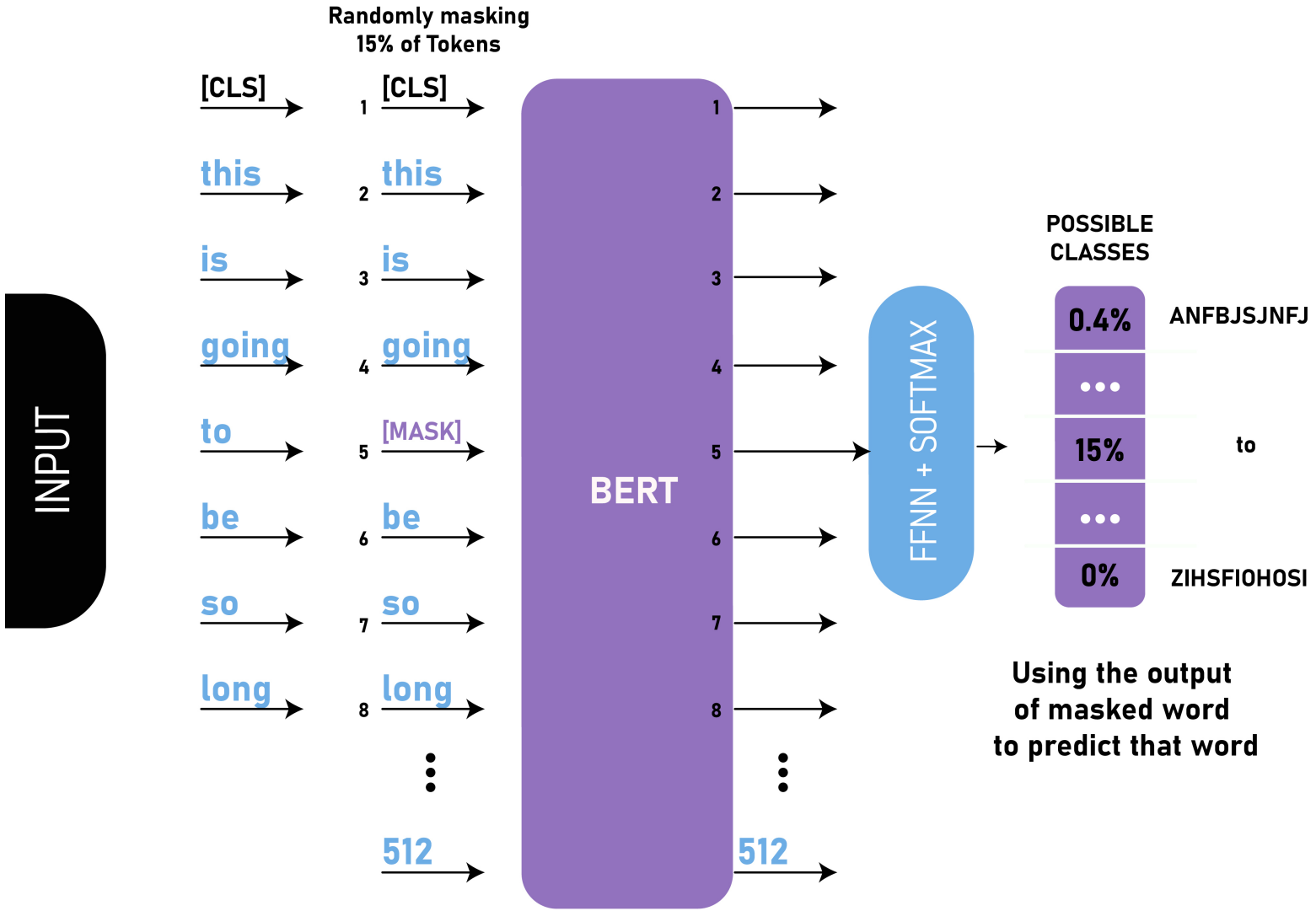
- To be good at the task, the context around the missing content has to be encoded. Therefore, it is necessary to learn semantic properties of an image.

Masked Language Models



- The same idea is used to train language models in NLP.
- Instead of masking parts of an image, we replace parts of the text with a special [MASK] token and predict the missing contents.
- Pretraining with massive amounts of text from the internet, leads to good performance on downstream tasks.
- In NLP, words/tokens are already separate semantic entities which makes the task generally better suited for masked predictions.

Masked Language Models



Prompting

Explaining a Joke

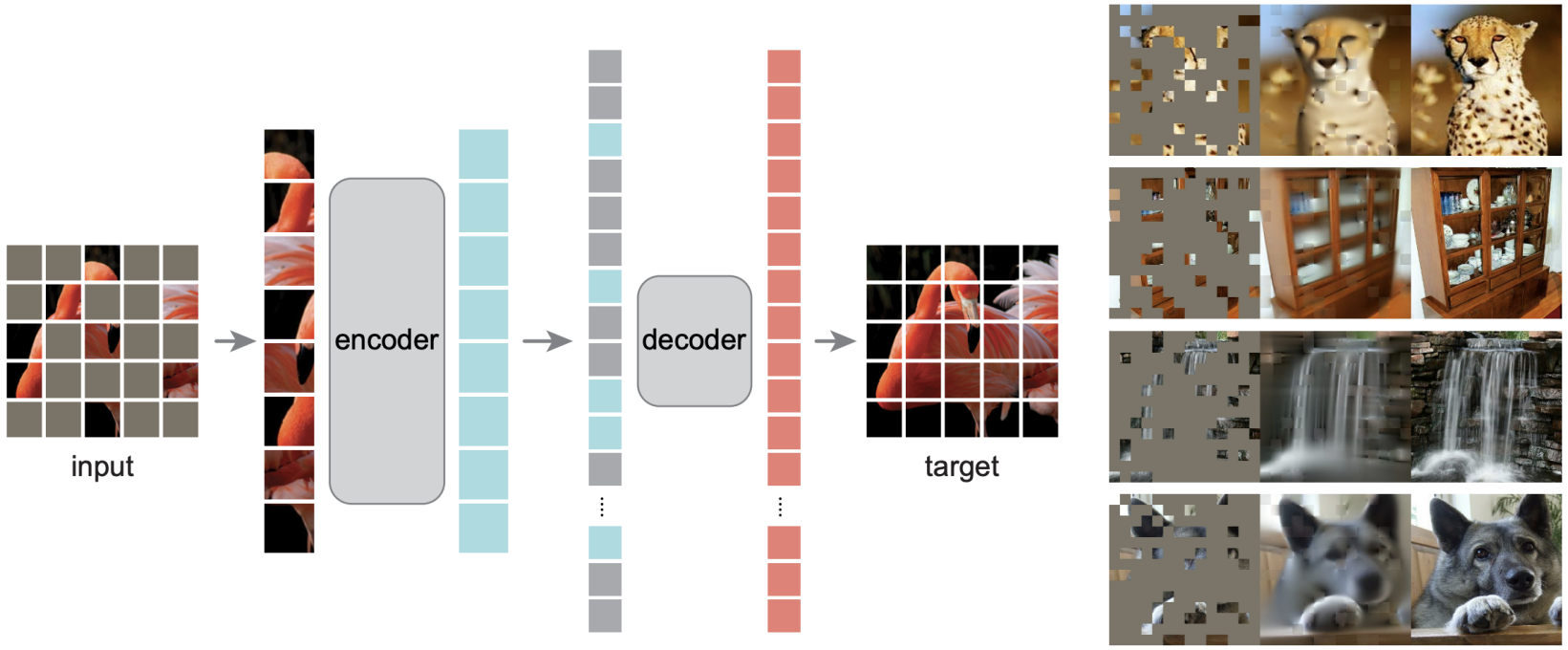
Input: Did you see that Google just hired an eloquent whale for their TPU team? It showed them how to communicate between two different pods!
Model Output: TPUs are a type of computer chip that Google uses for deep learning. A "pod" is a group of TPUs. A "pod" is also a group of whales. The joke is that the whale is able to communicate between two groups of whales, but the speaker is pretending that the whale is able to communicate between two groups of TPUs.

Logical Inference

Input: Shelley is from Virginia, but is visiting that city with that famous market where they throw the fish. Going home next Tuesday!
Question: Is it likely that Shelley will be near the Pacific Ocean this weekend?
Model Output: The city with the famous market where they throw the fish is Seattle, Washington. Seattle is on the Pacific Ocean. Shelley is visiting Seattle, so she will be near the Pacific Ocean this weekend. The answer is "yes", it is likely that Shelley will be near the Pacific Ocean this weekend.

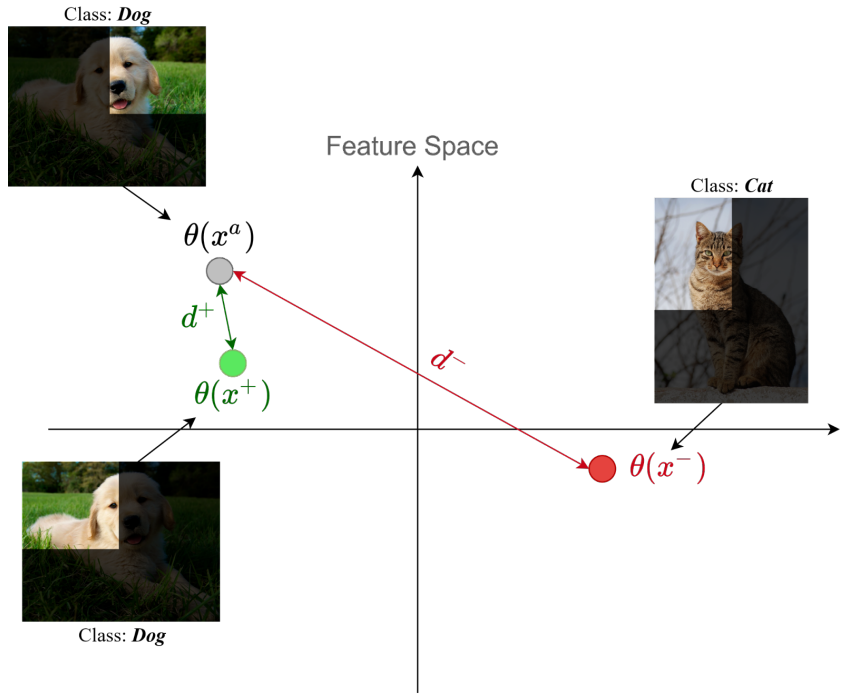
- This masked text pretraining task leads to interesting zero-shot capabilities when scaling training data and model size.
- We do not need any finetuning stage after the pretraining.
- By *prompting* the model with an example task and letting it predict the missing part of the sentence, the model is able to explain jokes or perform logical inference. [3].

Masked Autoencoders



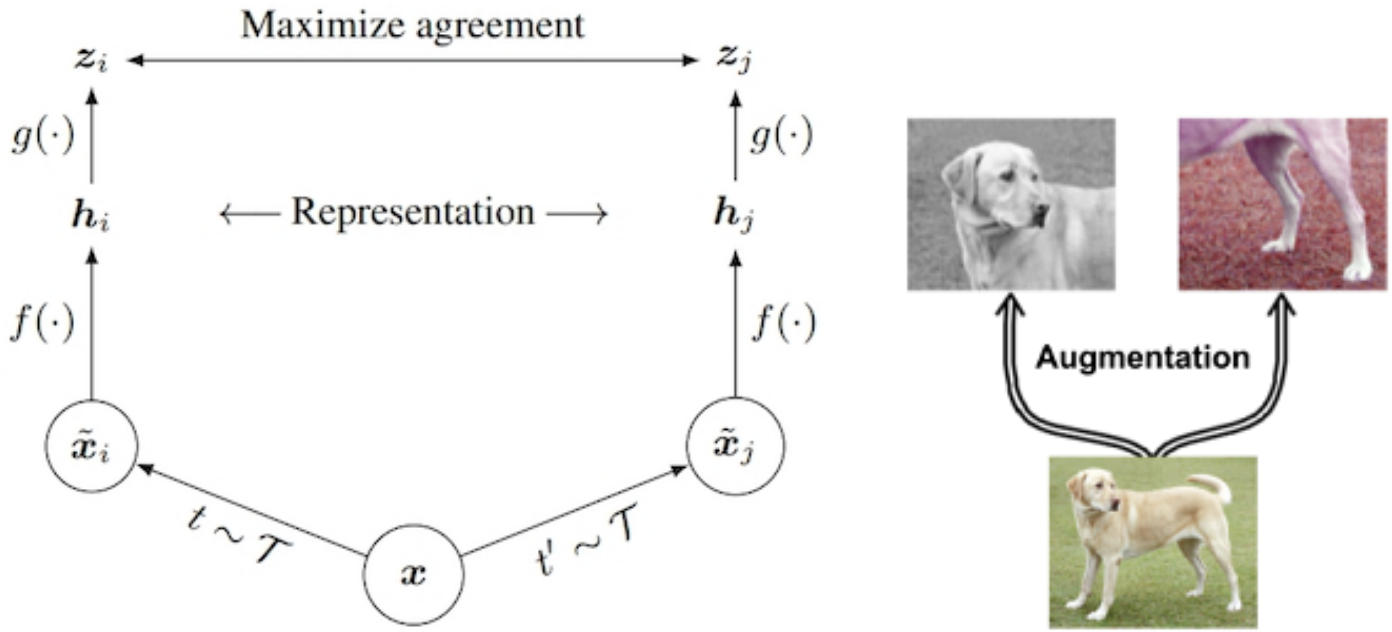
- A recent paper showed that mask prediction + the transformer architecture also works really well in vision.
- They discovered that a high masking ratio (75%) improves the representation quality.
- *Masked Autoencoders Are Scalable Vision Learners [6].*

Contrastive Learning



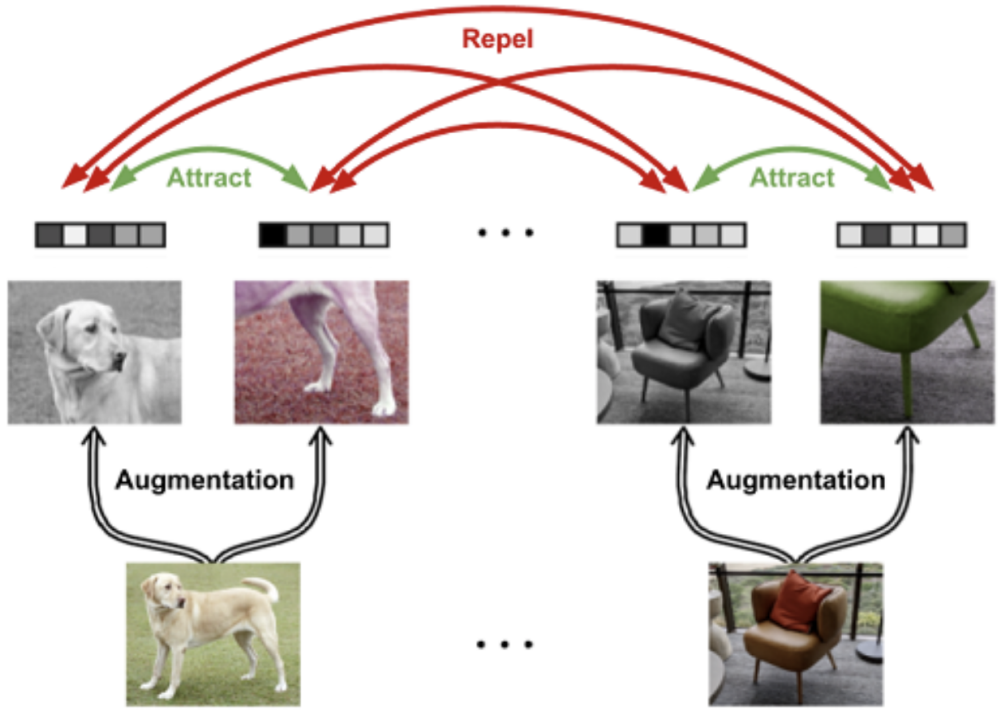
- In the previous pretext tasks, we designed the pretext task to be independent for each sample.
- Instead of looking at each sample individually, contrastive learning learns an embedding space in which similar sample pairs stay close to each other while dissimilar ones are far apart.
- We want to enforce that representations showing the same object are similar and dissimilar to representations from different objects.

SimCLR



- Simple Framework for Contrastive Learning of Visual Representations (SimCLR) [1, 2]
- Given one image, we augment two views $\tilde{x}_i, \tilde{x}_j$ from the image x and encode it with our network f.
- After projecting the representation h to embeddings z_i, z_j, we maximize the similarity between both embeddings.
- Problem: The easiest solution for the model would be to map all samples to the same constant vector. This is called **representation collapse**.

Negatives



- Therefore, we use other samples as **negatives** to prevent representation collapse.
- The negatives push the representation space apart to prevent the collapse to one constant point.
- SimCLR uses the other samples inside a batch as negatives.

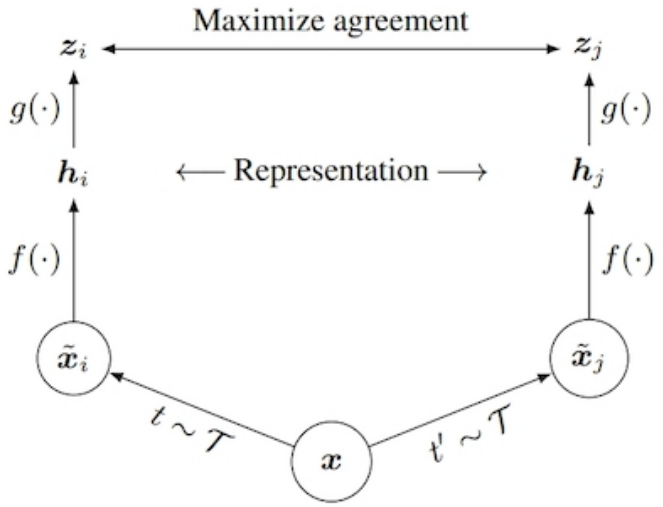
NT-Xent Loss

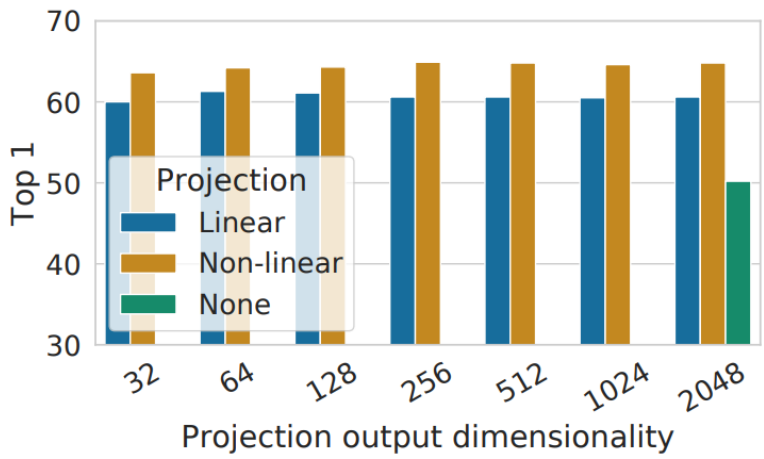
$$\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} [l(2k-1, 2k) + l(2k, 2k-1)] \tag{1}$$

$$l_{i,j} = -\log \frac{\exp\left(\text{sim}\left(\mathbf{z}_i, \mathbf{z}_j\right)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp\left(\text{sim}\left(\mathbf{z}_i, \mathbf{z}_k\right)/\tau\right)} \tag{2}$$

- We maximize the similarity between representations of views z_i, z_j from the same image.
- and minimize similarity between views z_i, z_k coming from different views inside the batch.
- Cosine similarity is usually used as a similarity measure $sim(u, v) = \frac{u \cdot v}{||u|| \cdot ||v||}$.

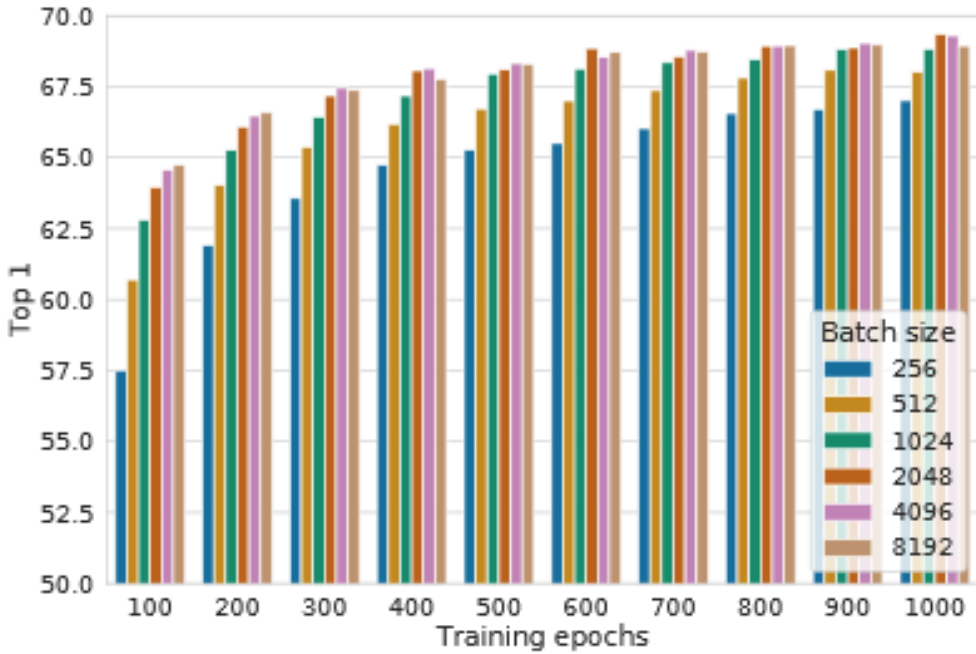
Projection Head





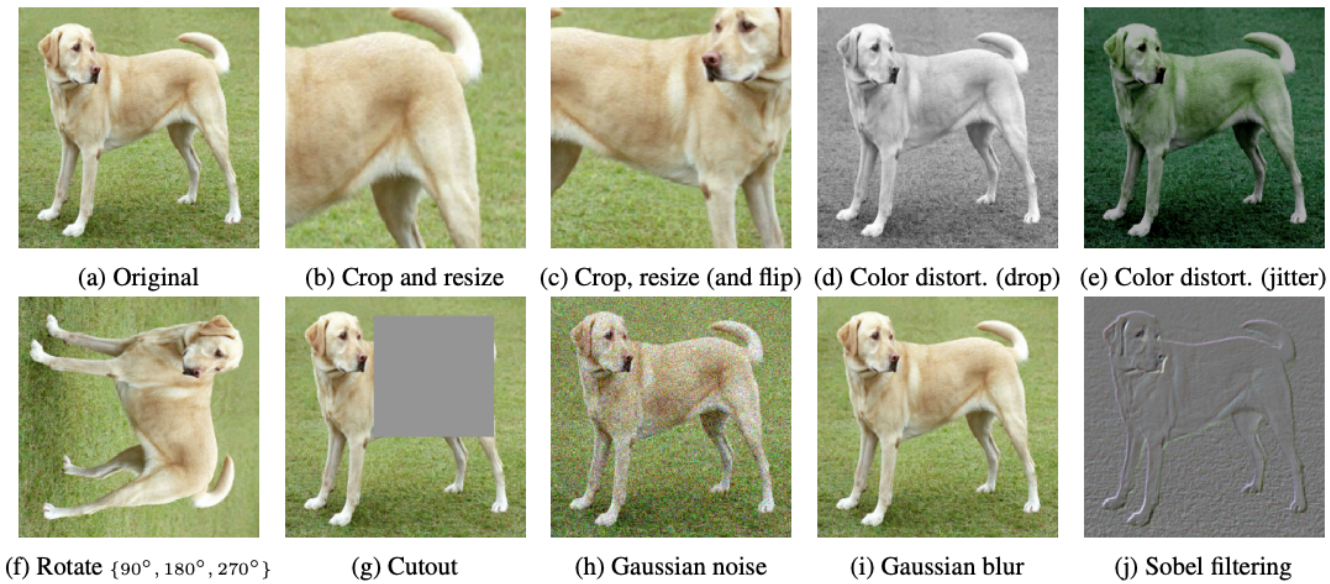
- The representations h_i are used for linear probing in the downstream task.
- The projection head is discarded after the pretraining stage.
- A non-linear projection head g improves linear evaluation performance.

Large Batch Size



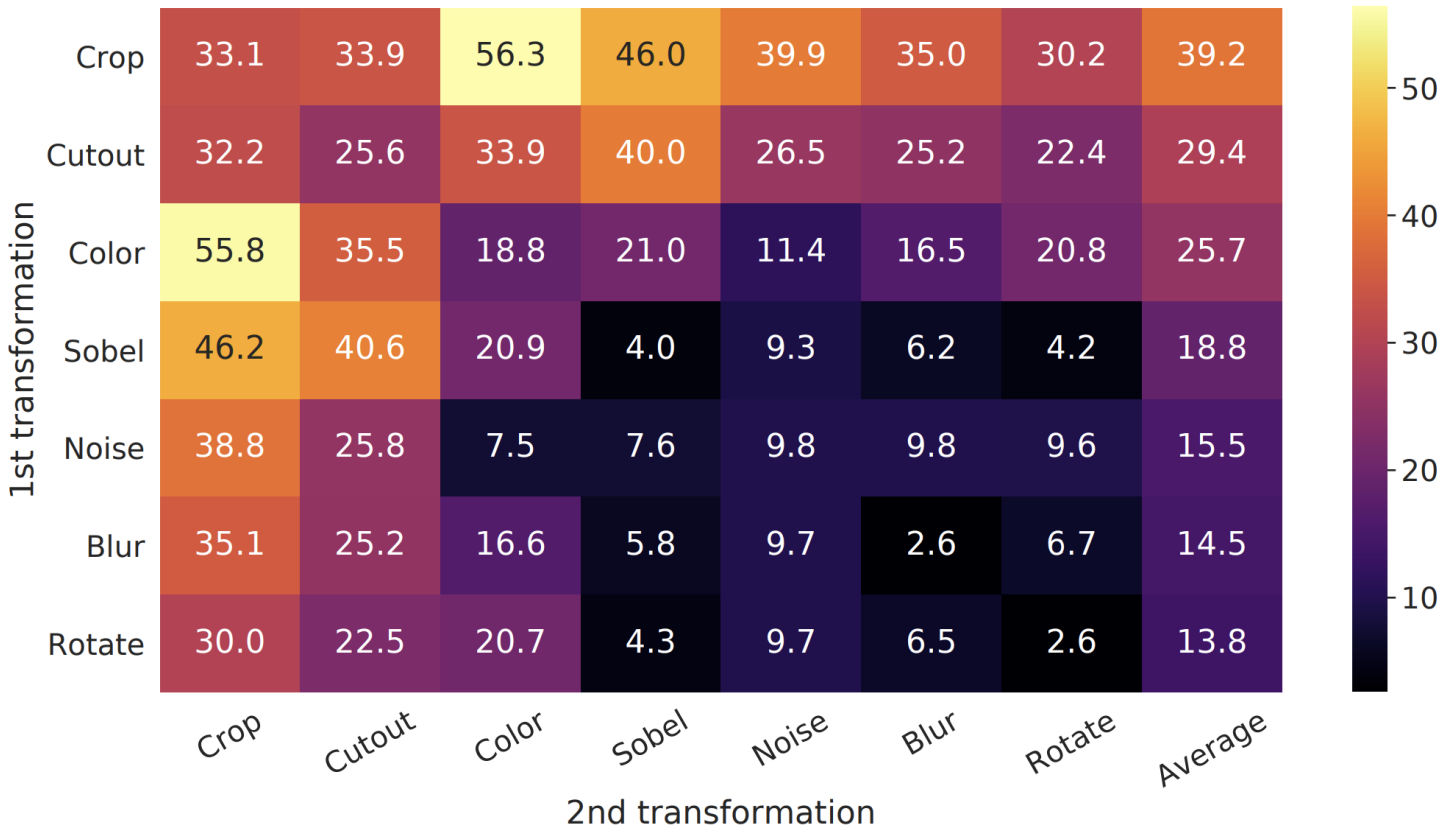
- SimCLR needs a lot of negatives samples in one batch to learn good representations (batch size 4096).
- Therefore, training with multiple GPUs or TPUs is necessary.
- Other contrastive learning approaches (e.g. MoCo [7]) circumvent this by storing representations in a memory bank and use them as negatives in future iterations.

Augmentations



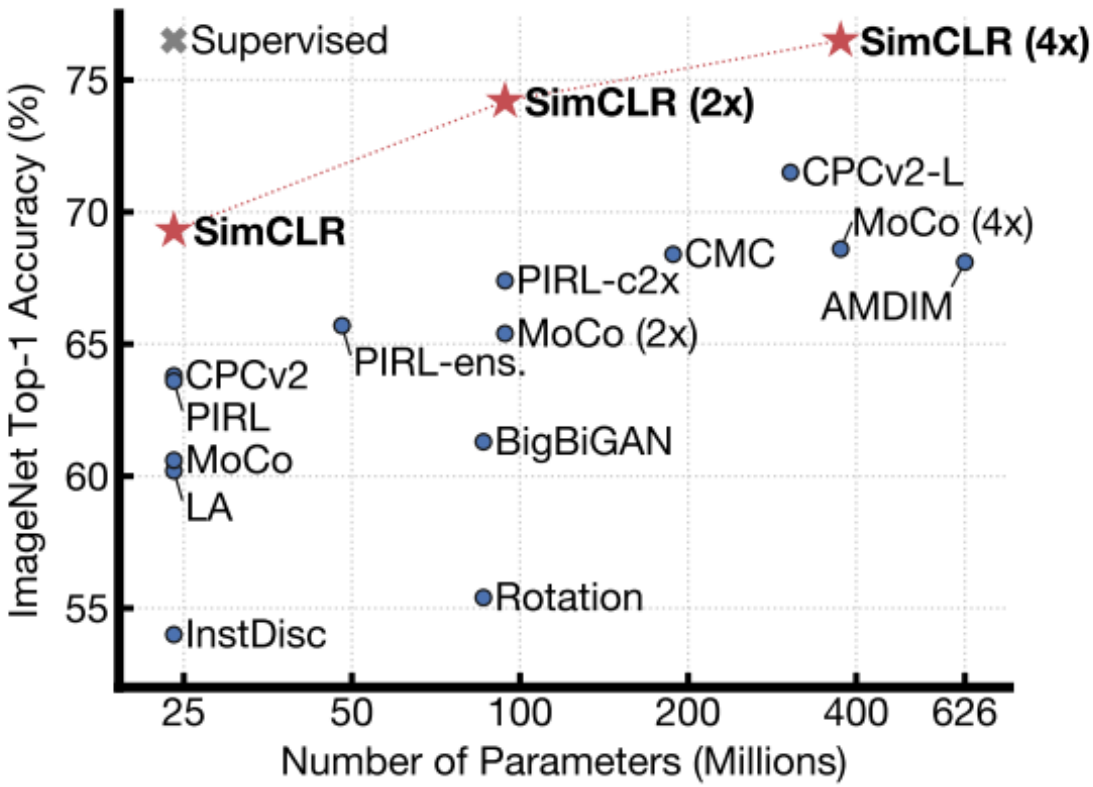
- SimCLR sequentially applies three simple augmentations: random cropping followed by resize back to the original size, random color distortions, and random Gaussian blur.
- Augmentations are tuned towards performance on ImageNet.
- For other types of images or domains (e.g. histopathology images, text), we have to handcraft the appropriate type of augmentations.

Augmentation Grid Search



- The authors performed a grid search on ImageNet and the combination of color jittering and cropping is crucial for good performance.

Performance



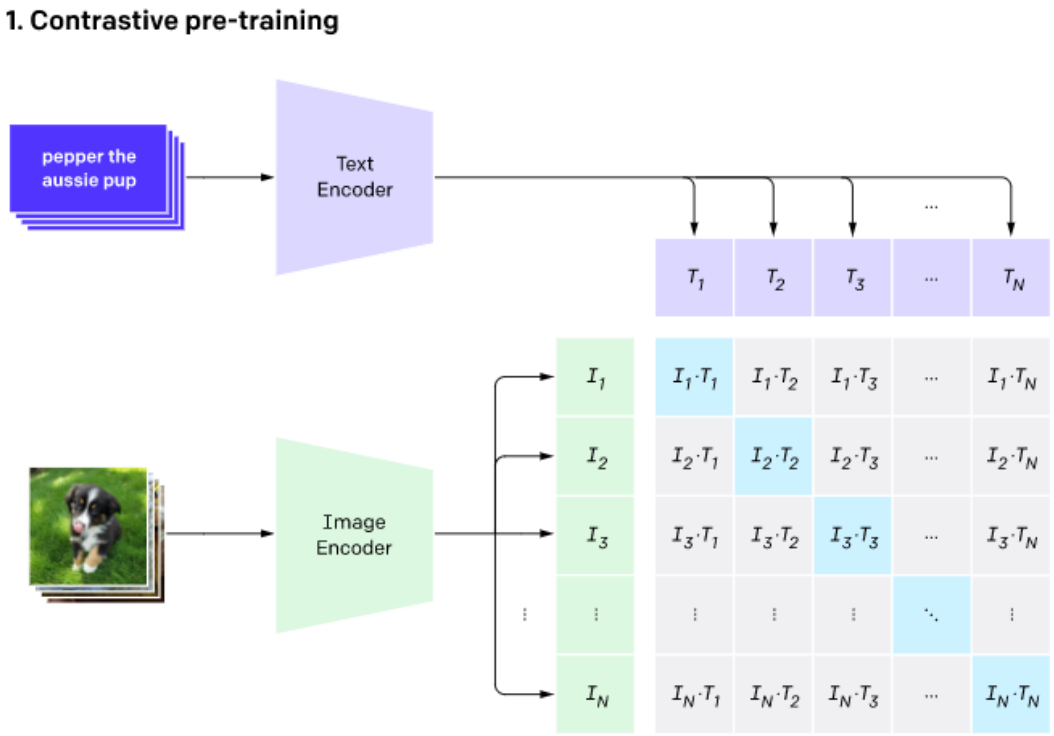
- Linear probing accuracy challenges supervised ImageNet performance.
- Newer methods even further close the gap.

Transfer Learning

	Food	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech-101	Flowers
Linear evaluation:												
SimCLR (ours)	**76.9**	**95.3**	80.2	48.4	**65.9**	60.0	61.2	**84.2**	**78.9**	89.2	**93.9**	**95.0**
Supervised	75.2	**95.7**	**81.2**	**56.4**	64.9	**68.8**	**63.8**	83.8	**78.7**	**92.3**	**94.1**	94.2
Fine-tuned:												
SimCLR (ours)	**89.4**	**98.6**	**89.0**	**78.2**	**68.1**	**92.1**	**87.0**	**86.6**	**77.8**	92.1	**94.1**	97.6
Supervised	88.7	98.3	**88.7**	**77.8**	67.0	91.4	**88.0**	86.5	**78.8**	**93.2**	**94.2**	**98.0**
Random init	88.3	96.0	81.9	**77.0**	53.7	91.3	84.8	69.4	64.1	82.7	72.5	92.5

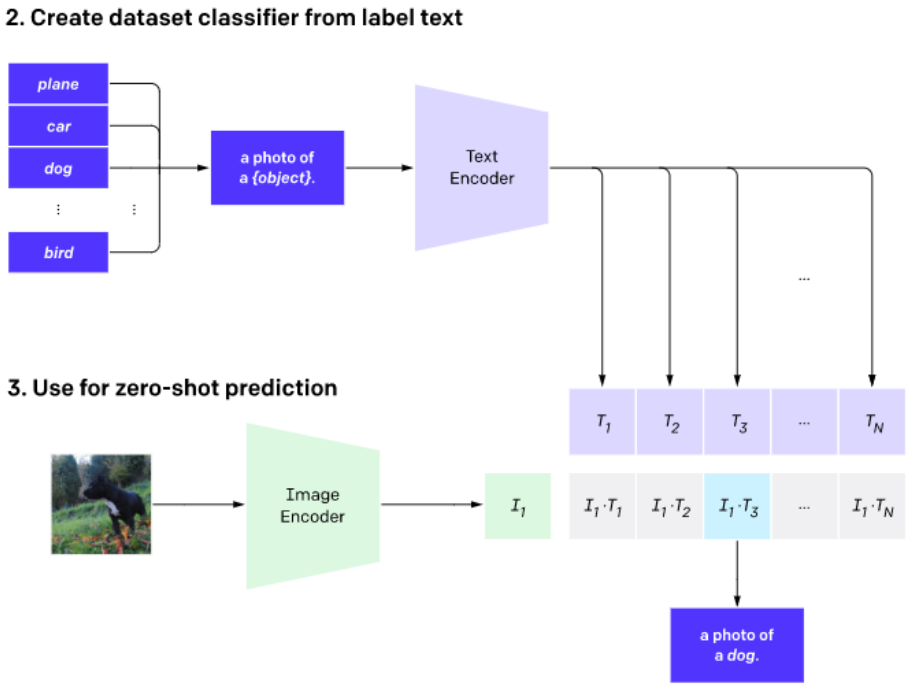
- Transfer Learning to other downstream datasets works better from self-supervised models compared to supervised pretraining.
- This is crucial as an ImageNet initialized model is mostly used for transfer learning to a different downstream task.
- In domains where there is not an ImageNet scale labeled dataset available (e.g. histopathology), this drastically improves downstream task performance in many applications.

Image-Text Models



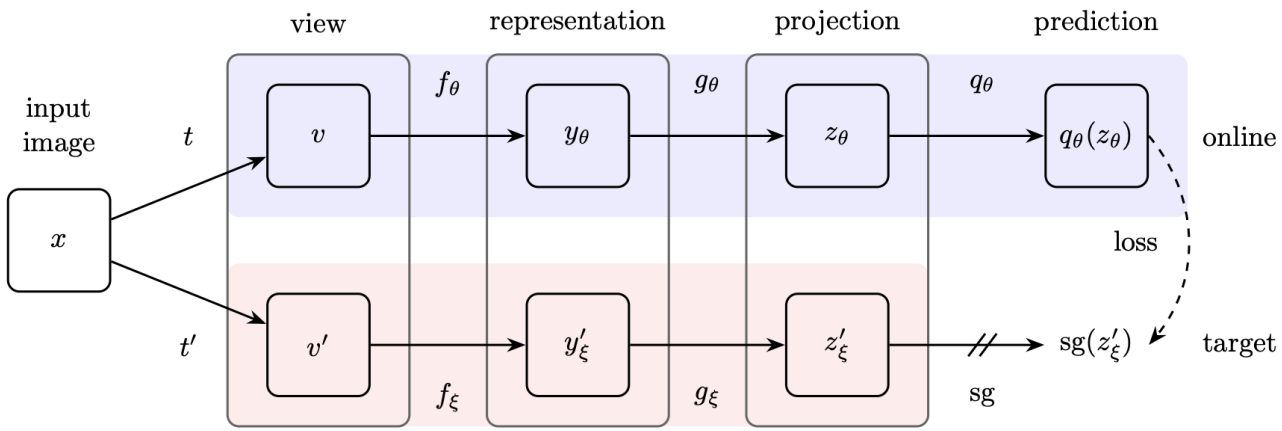
- Use image-text pairs for contrastive learning (e.g. CLIP [10])
- Instead of two views, we maximize the similarity between an image and its corresponding caption.
- Instead of using the same encoder for both views, we have a text encoder (transformer model) and an image encoder (CNN, ViT).

Image-Text Models



- The caption/image pairs are scraped from the internet.
- OpenAI pretrained CLIP with 400M image/text pairs. There are now also open-source datasets available with 5B image/text pairs (Laion-5B).
- Mapping images and text into the same representation space allows more applications that have not been possible before.
  - Image generation models (DALL-E, Stable Diffusion, Midjourney).
  - zero-shot image classification
  - text $\rightarrow$ image search

Non-Contrastive Learning



- Non-Contrastive learning also uses two augmented views from the same image, but prevents representation collapse without using other samples as negatives.
- Bootstrap your own latent (BYOL) [5].
- predict the representations of a random initialized model improves the representations.
- BYOL uses this insight iteratively and always predicts the representations of a momentum encoded version of the encoder.

iBOT and DINOv2: Joint Predictions with Masking

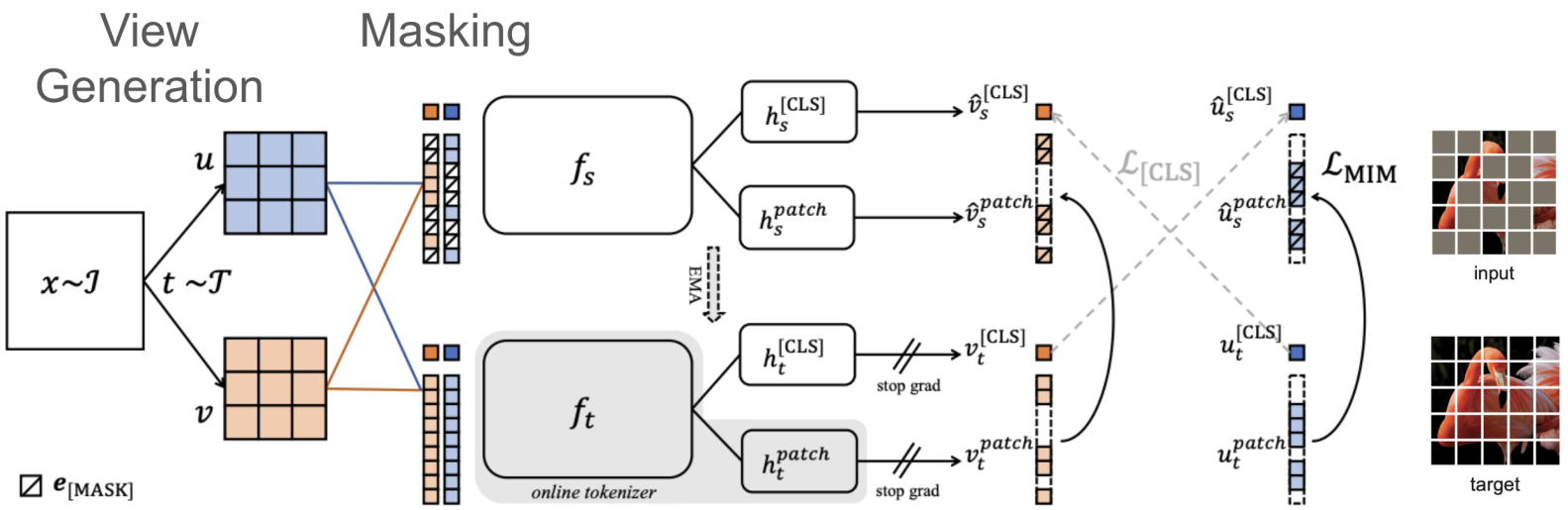
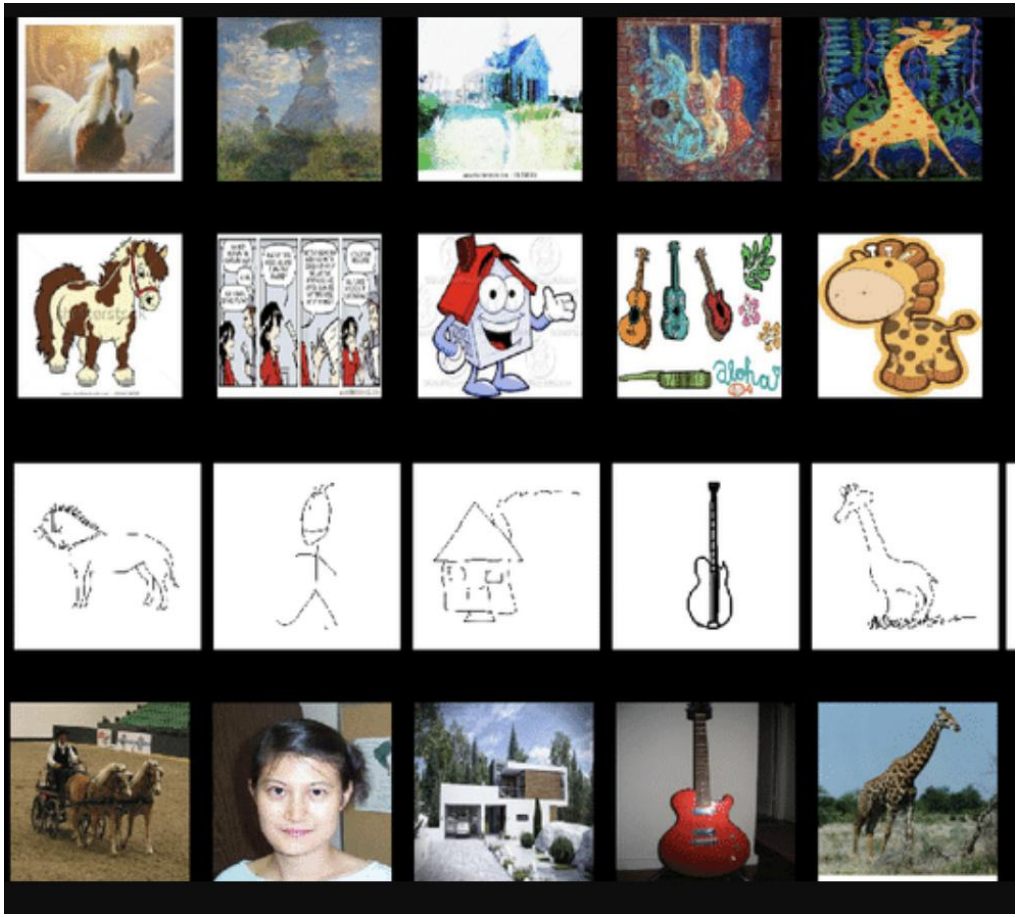


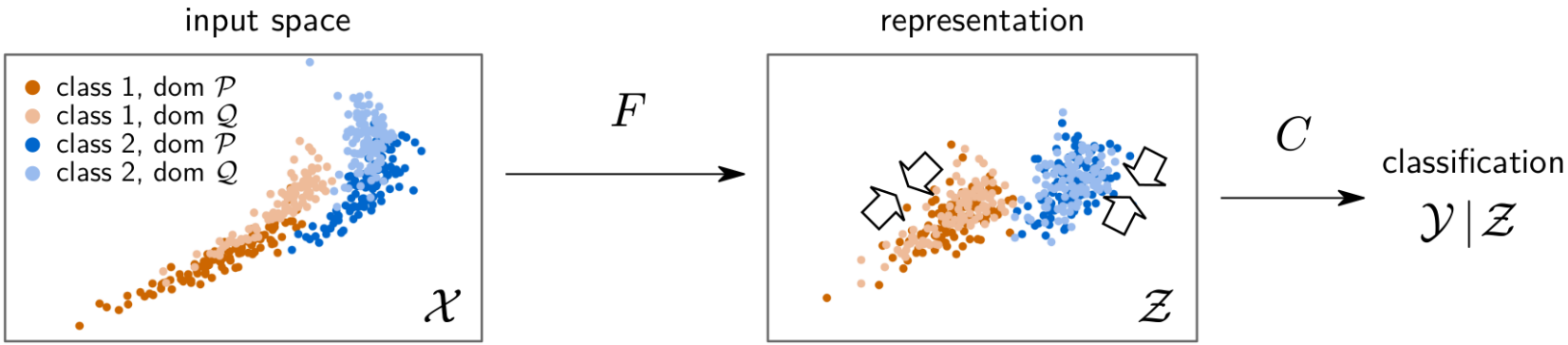
Figure 3: **Overview of iBOT framework, performing masked image modeling with an *online tokenizer*.** Given two views $\boldsymbol{u}$ and $\boldsymbol{v}$ of an image $\boldsymbol{x}$, each view is passed through a teacher network $h_t \circ f_t$ and a student network $h_s \circ f_s$. iBOT minimizes two losses. The first loss $\mathcal{L}_{\texttt{[CLS]}}$ is self-distillation between cross-view `[CLS]` tokens. The second loss $\mathcal{L}_{\text{MIM}}$ is self-distillation between in-view patch tokens, with some tokens masked and replaced by $\boldsymbol{e}_{\texttt{[MASK]}}$ for the student network. The objective is to reconstruct the masked tokens with the teacher networks' outputs as supervision.

Learning Domain Invariant Representations

In practice, data often comes from multiple domains. We want a model that predicts well for each domain (even those that are the most difficult or that have fewer labels).



Learning Domain Invariant Representations



- *Approach:* learn a function F that maps the data to a representation where the domains cannot be differentiated, and from which a domain-invariant classifier C can be built.
- Domain invariance can be measured by how low the *Wasserstein distance* between distributions of to each domain are.
- *The Wasserstein distance $W(\mathcal{P}, \mathcal{Q})$ measures how much it costs (in Euclidean sense) to transport every bit of distribution $\mathcal{P}$ onto $\mathcal{Q}$.*

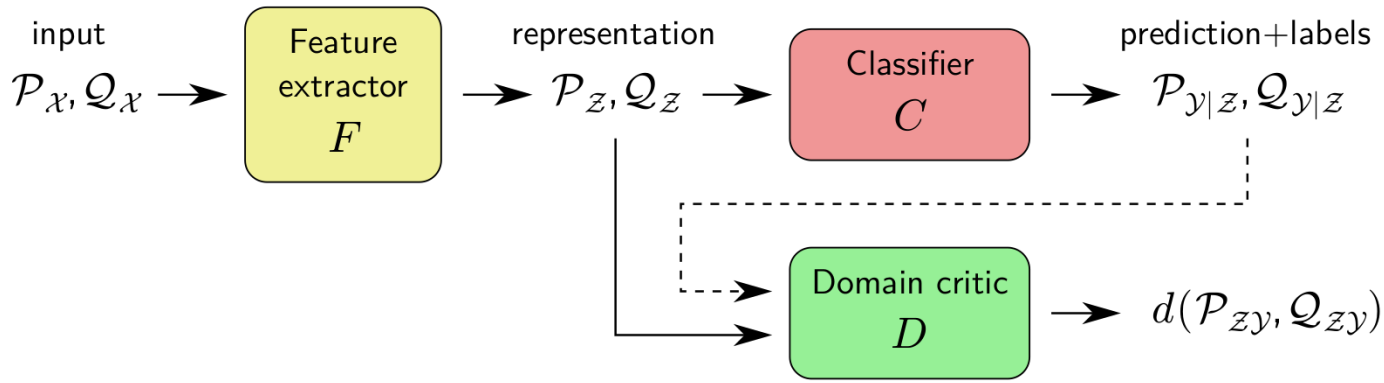
Reference publication: Andéol et al. (2021) Learning Domain Invariant Representations by Joint Wasserstein Distance Minimization

Learning Domain Invariant Representations

Insight: The Wasserstein distance (on the joint representation-label space) between the two domains $\mathcal{P}$ and $\mathcal{Q}$, i.e.

$$W(\mathcal{P}_{\mathcal{Z}\mathcal{Y}}, \mathcal{Q}_{\mathcal{Z}\mathcal{Y}})$$

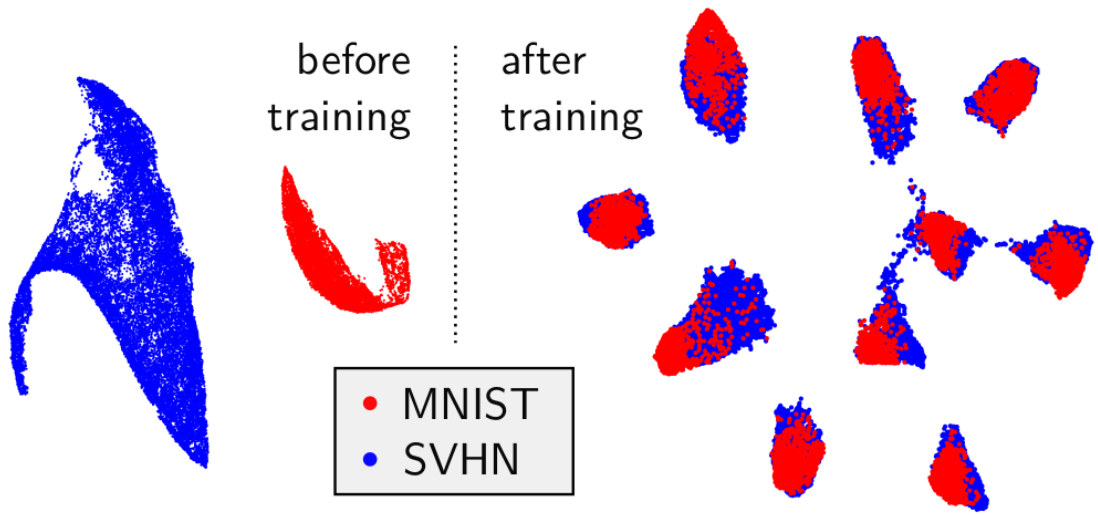
can be upper-bounded in a way that it becomes expressible as combination of common and easily optimizeable neural network modules (classifier, domain critic, etc.):



In this architecture, domain invariance is achieved by simultaneously (1) learning to detect the domain from the representation (domain critic D), and (2) learning a feature extractor F that *fools* the domain critic D.

Learning Domain Invariant Representations

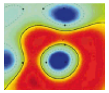
The process of learning a domain invariant representation can be visualized by applying UMAP (a low-dimensional embedding method similar to T-SNE) in the representation space, and color-coding points by their domain.



Before training, the representation is strongly dominated by domain membership (MNIST: black&white handwritten digits, SVHN: color printed digits). After training, distributions are becoming aligned (hard to distinguish between domains) and form clusters corresponding to the different classes.

Summary

- Representation learning has the goal to learn semantic vector representations for high-dimensional data in an unsupervised fashion.
- The self-supervised pretraining objective is called **Pretext Task**.
- **Contrastive Learning** learns aims to learn an embedding space in which similar sample pairs stay close to each other while dissimilar ones are far apart.
- SimCLR:
  - uses negatives to prevent representation collapse.
  - uses large batch sizes.
  - a non-linear projection head.
- We can also learn similarity/dissimilarity between different data modalities (e.g. image/text pairs).
- We can explicitly learn representations with other properties that suit the downstream task (e.g. domain invariance).



References I

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton.
A simple framework for contrastive learning of visual representations.
In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton.
Big self-supervised models are strong semi-supervised learners.
Advances in neural information processing systems, 33:22243–22255, 2020.

A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al.
Palm: Scaling language modeling with pathways.
arXiv preprint arXiv:2204.02311, 2022.

S. Gidaris, P. Singh, and N. Komodakis.
Unsupervised representation learning by predicting image rotations.
arXiv preprint arXiv:1803.07728, 2018.

J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al.
Bootstrap your own latent-a new approach to self-supervised learning.
Advances in neural information processing systems, 33:21271–21284, 2020.

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick.
Masked autoencoders are scalable vision learners.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick.
Momentum contrast for unsupervised visual representation learning.
In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

References II

M. Noroozi and P. Favaro.

Unsupervised learning of visual representations by solving jigsaw puzzles.
In *European conference on computer vision*, pages 69–84. Springer, 2016.

D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros.

Context encoders: Feature learning by inpainting.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al.

Learning transferable visual models from natural language supervision.
In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

R. Zhang, P. Isola, and A. A. Efros.

Colorful image colorization.
In *European conference on computer vision*, pages 649–666. Springer, 2016.