PHAROS

THE GREEK AI FACTORY

AI4Health Track

Training Series

Course 6

Vision Representation Learning and Generative Models in Biomedicine

MARCH 16, 2026 | 12:00 EET | ONLINE

EuroHPC Joint Undertaking

Co-funded by the European Union

PHAROS GREECE

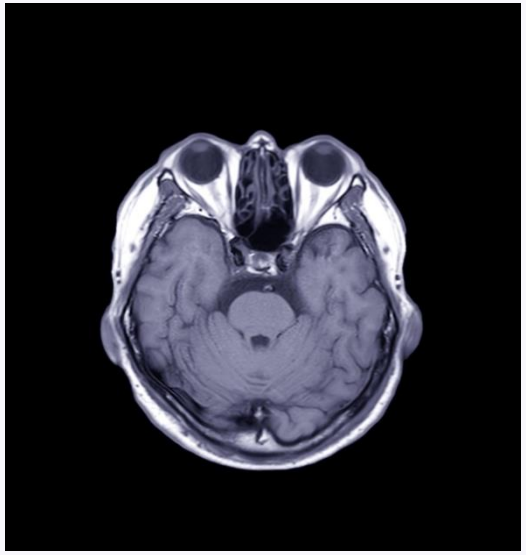
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Ψηφιακής Διακυβέρνησης
και Τεχνητής Νοημοσύνης

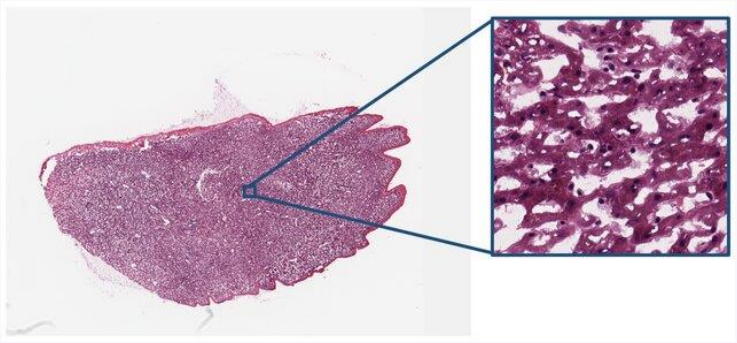
Course Agenda

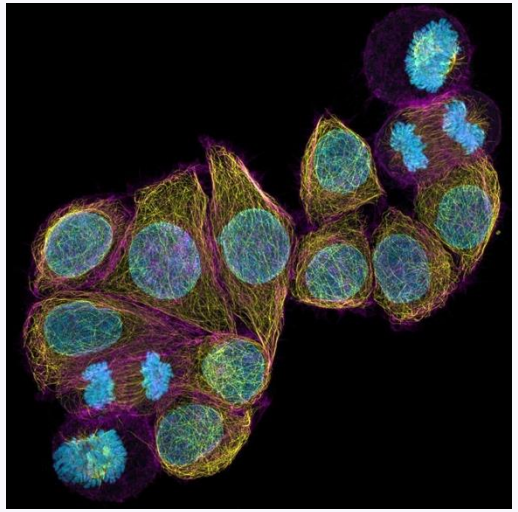
Part I — Introduction & Objectives
Part II — Vision Representation Learning
Part III — Self-Supervised Learning
Part IV — Generative Models
Part V — Emerging Directions & Discussion

Part I — Introduction & Objectives

Biomedical Imaging is Exploding

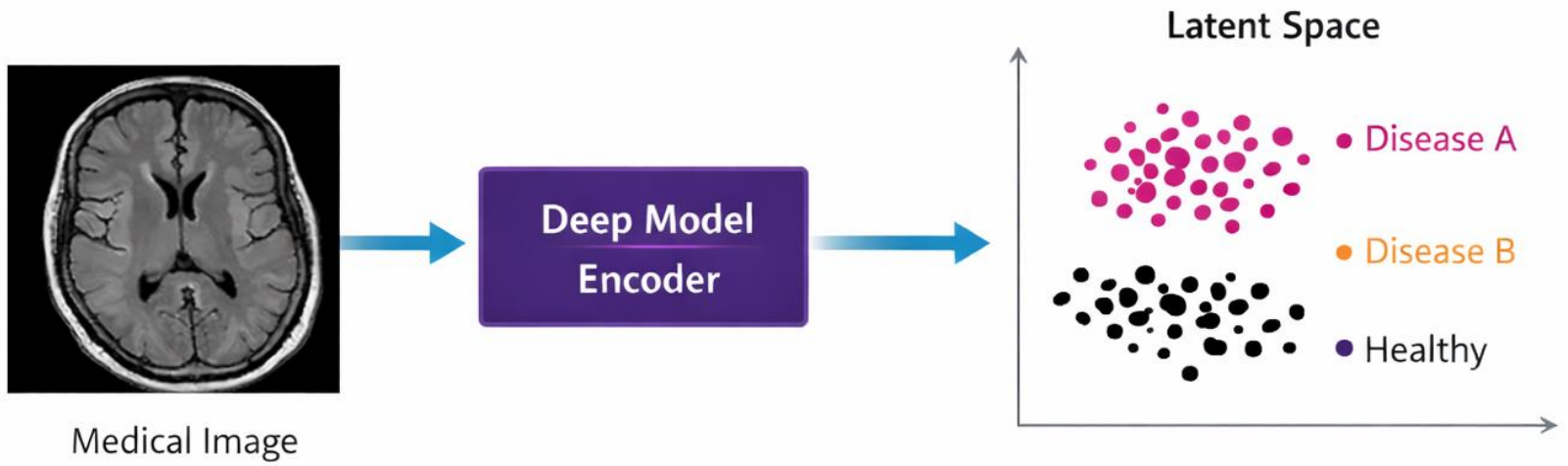






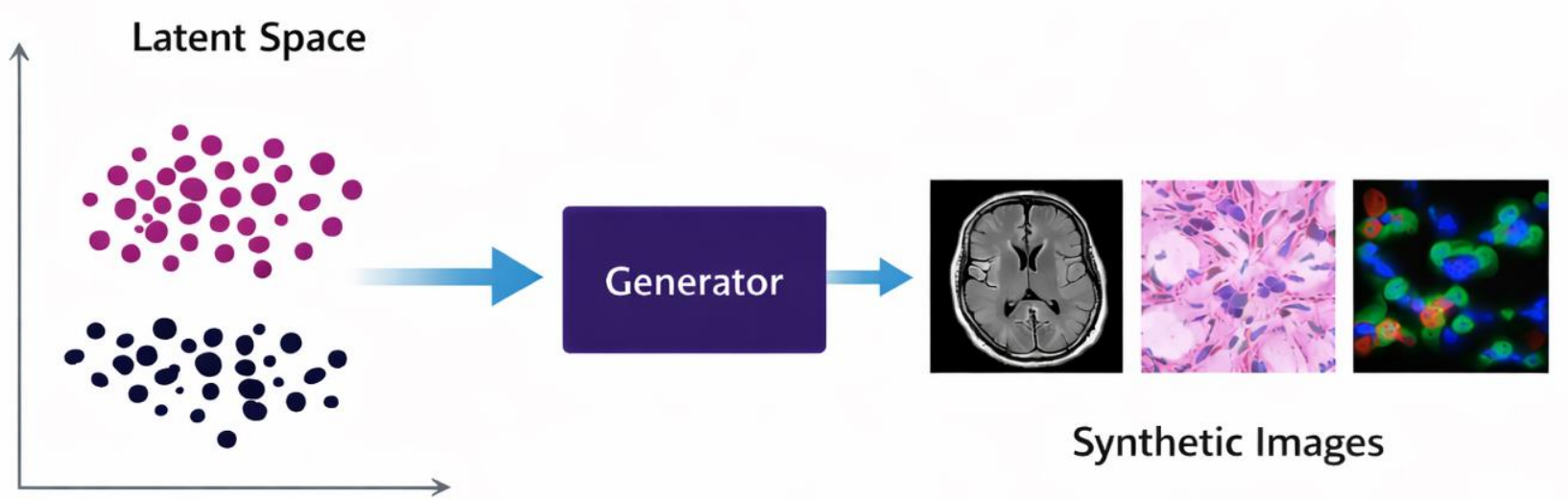
- Massive image data across radiology, pathology, microscopy
- Limited expert labels
- High clinical stakes
- Hidden biological structure

From Pixels to Representations



- Representation learning: map images → meaningful features
- Latent space: compact description of image structure
- Goal: capture biologically relevant variation

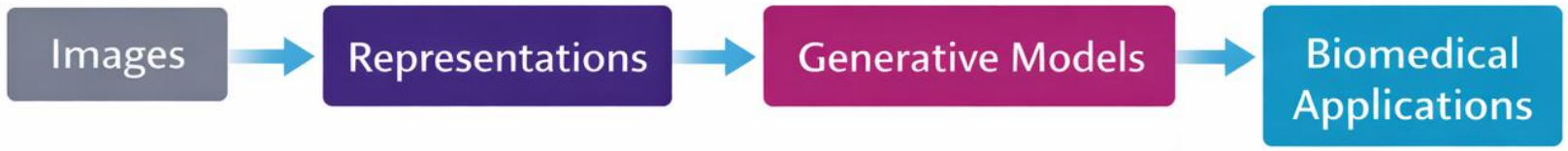
From Representation to Generation



- Understanding structure → learn latent space
- Sampling from latent space → generate new images
- Applications → simulation, augmentation, modality translation

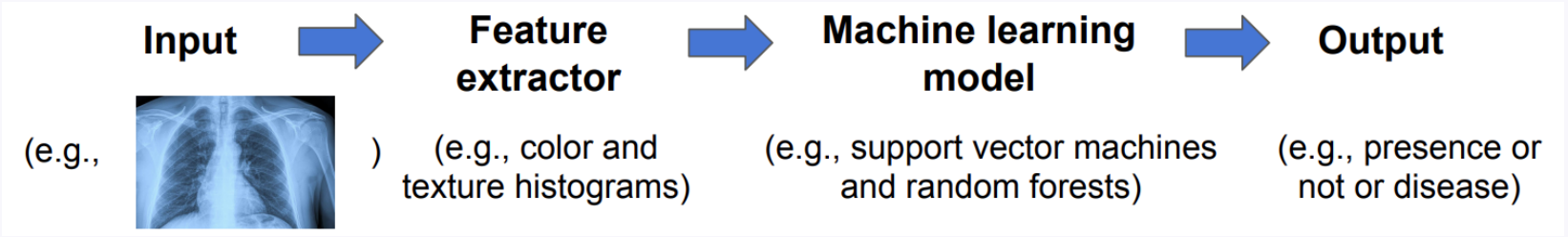
What We Will Cover

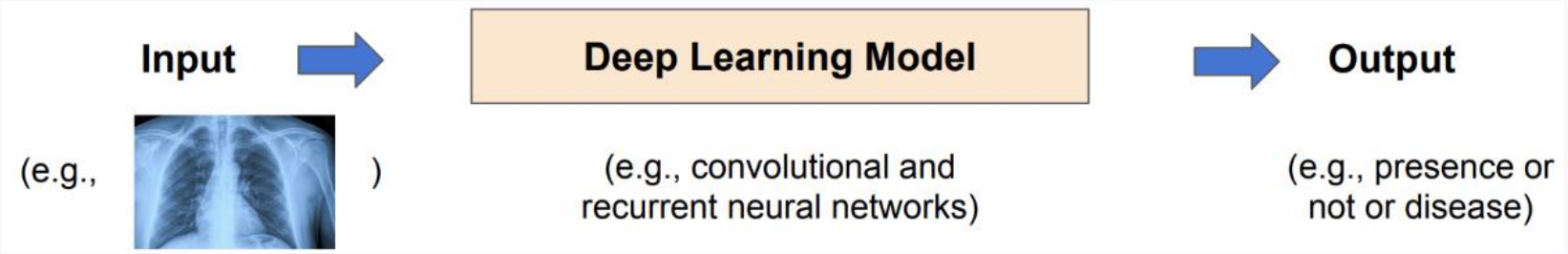
- **Representation learning** in biomedical imaging
- **Self-supervised learning** when labels are scarce
- **Generative models**: VAE, GAN, Diffusion
- **Future directions**



Part II — Vision Representation Learning

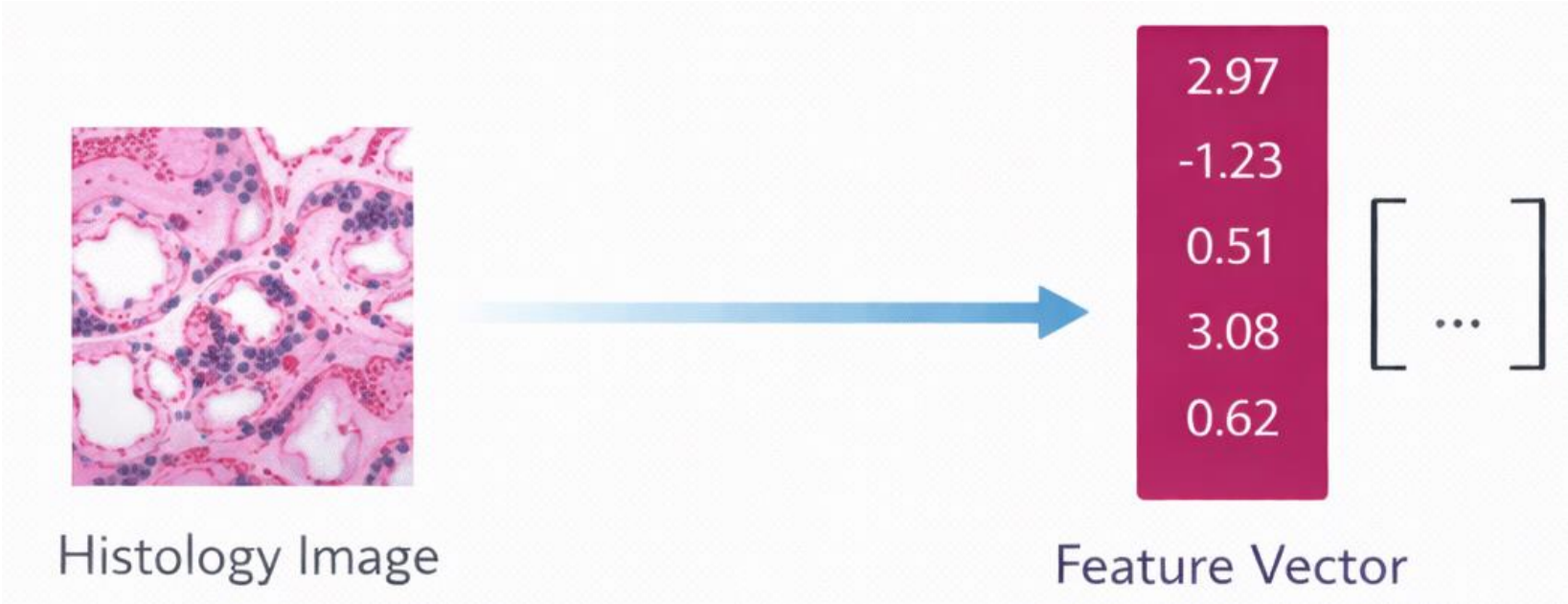
Why Representation Learning?





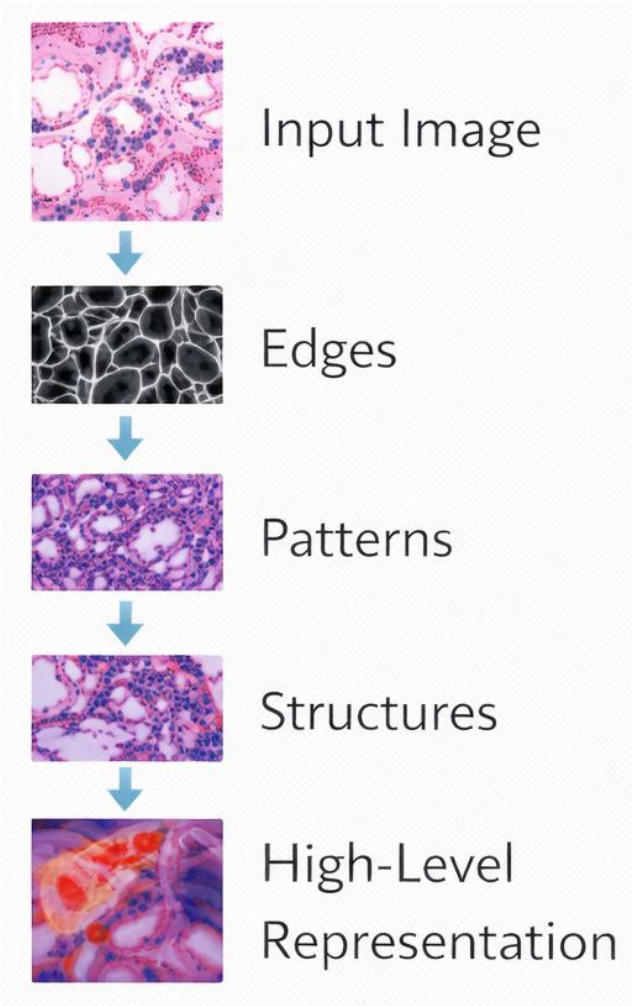
- Traditional computer vision relied on handcrafted features
- Deep Learning learns representations directly from data
- End-to-end learning improves performance and scalability

What is a Representation?



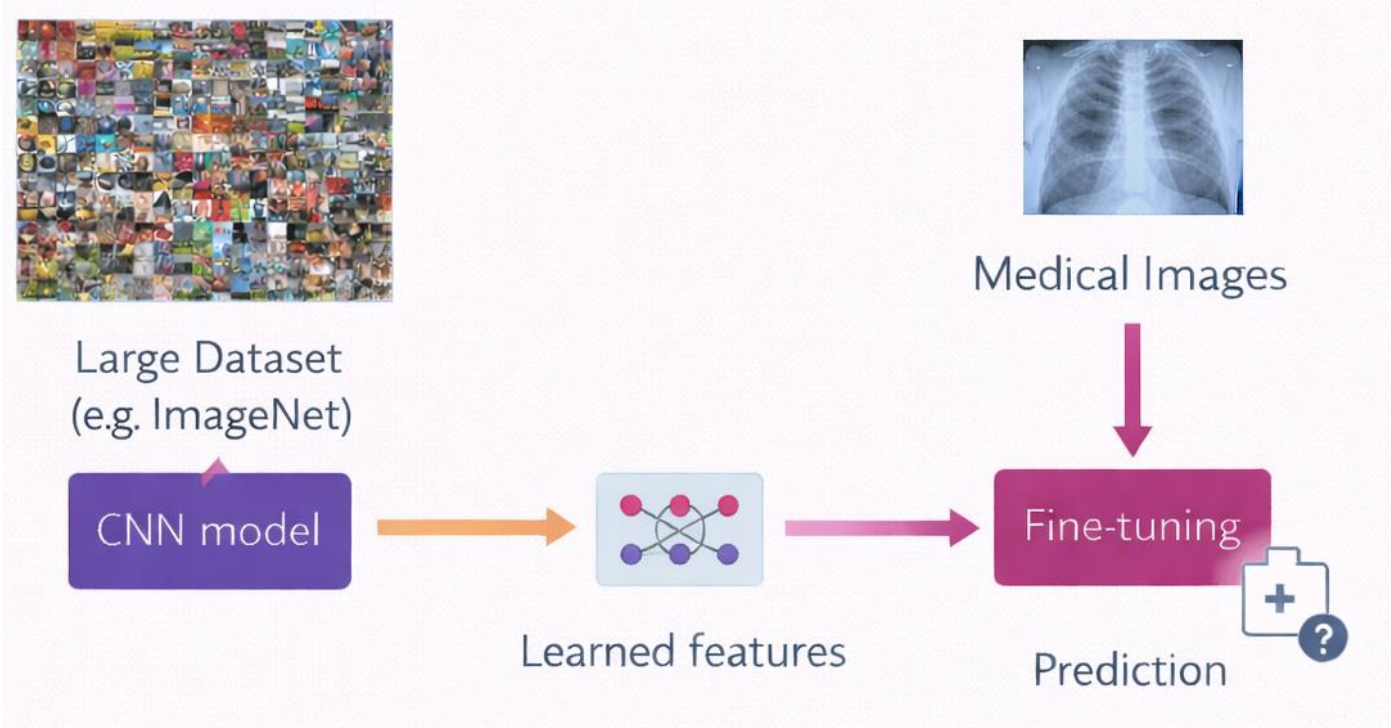
- **Compact encoding** → image summarized as a vector
- **Captures task-relevant information** → disease patterns, morphology
- **Reduces dimensionality** → millions of pixels → smaller feature space
- **Enables generalization** → similar images have similar representations

Deep Networks Learn Hierarchical Representations



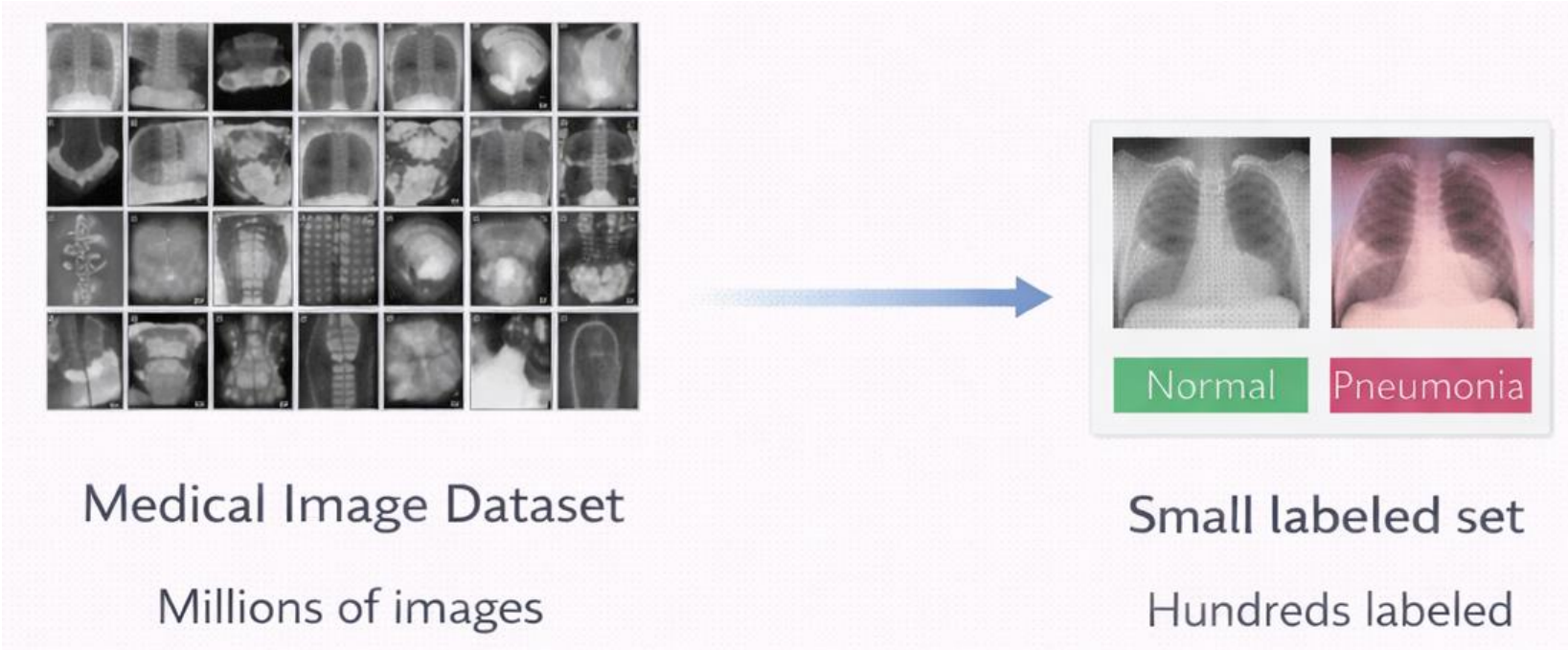
- **Early** layers detect simple patterns (edges, gradients)
- **Intermediate** layers detect textures and shapes
- **Deep** layers capture semantic structures

Transfer Learning in Medical Imaging



- **Pretraining** on large datasets learns general features
- **Fine-tuning** adapts the model to biomedical tasks
- Improves performance with limited labelled data

The Label Bottleneck in Biomedical Imaging

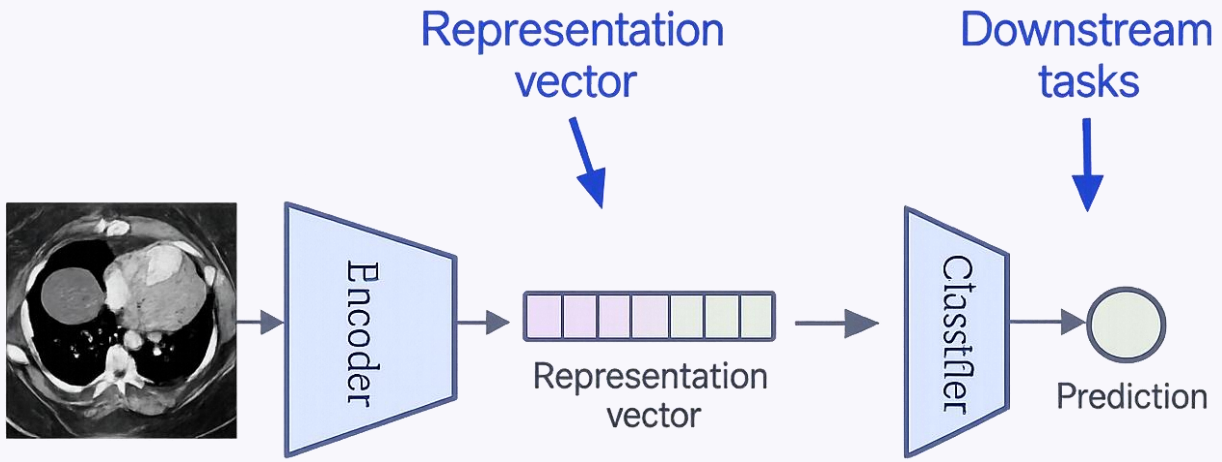


- Medical datasets are large but weakly labelled
- Expert annotation is expensive
- Supervised learning does not scale

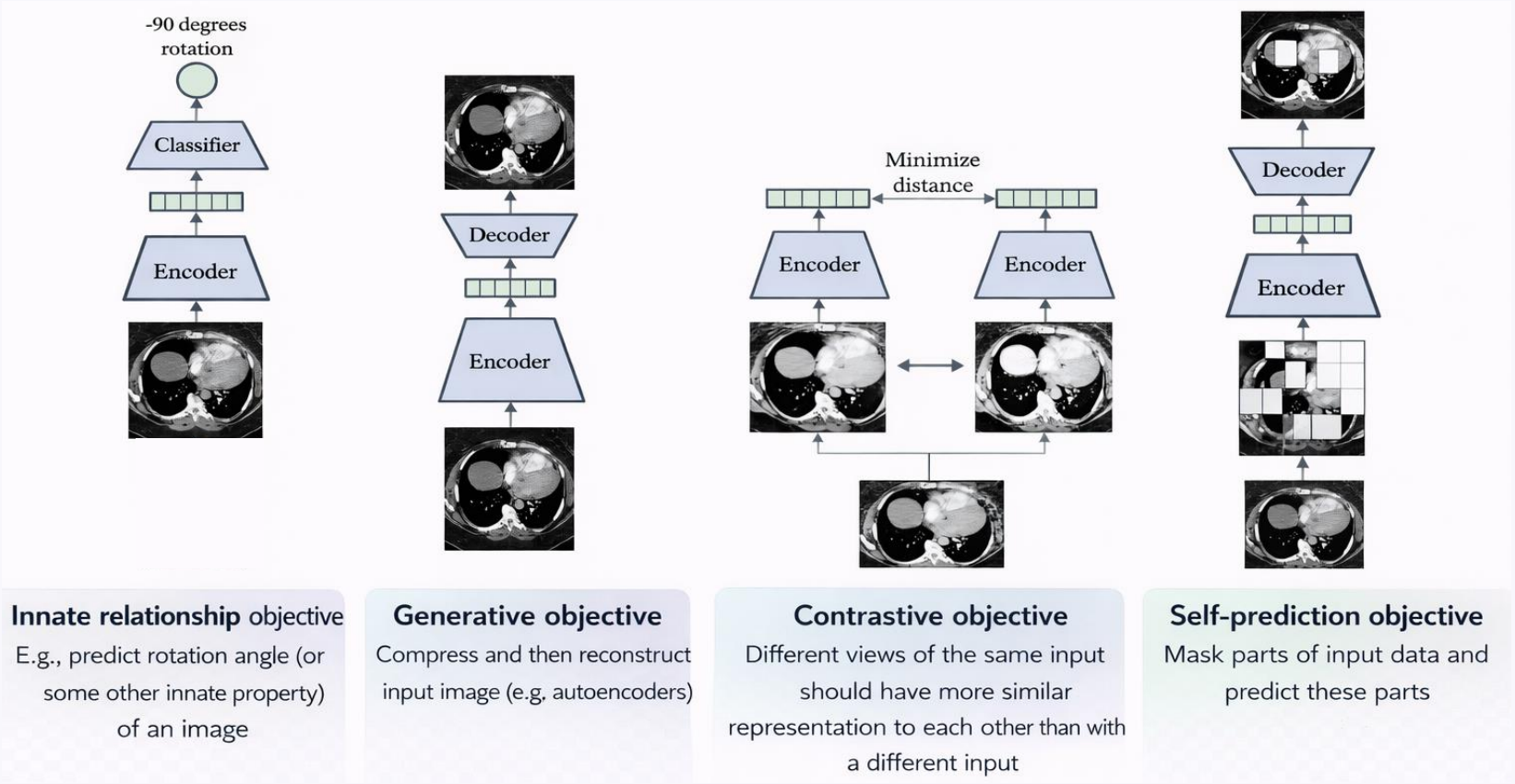
Part III — Self-Supervised Learning

Representation Learning for Images

- **Goal**: learn to extract low-dimensional feature **representations** from images that capture meaningful semantic information and can be reused across many downstream tasks.
- These representations are often called "**embeddings**", and the neural networks that produce them are commonly referred to as "***encoders***" or "**embedding models**"



Different representation learning paradigms



Innate relationship objective
E.g., predict rotation angle (or some other innate property) of an image

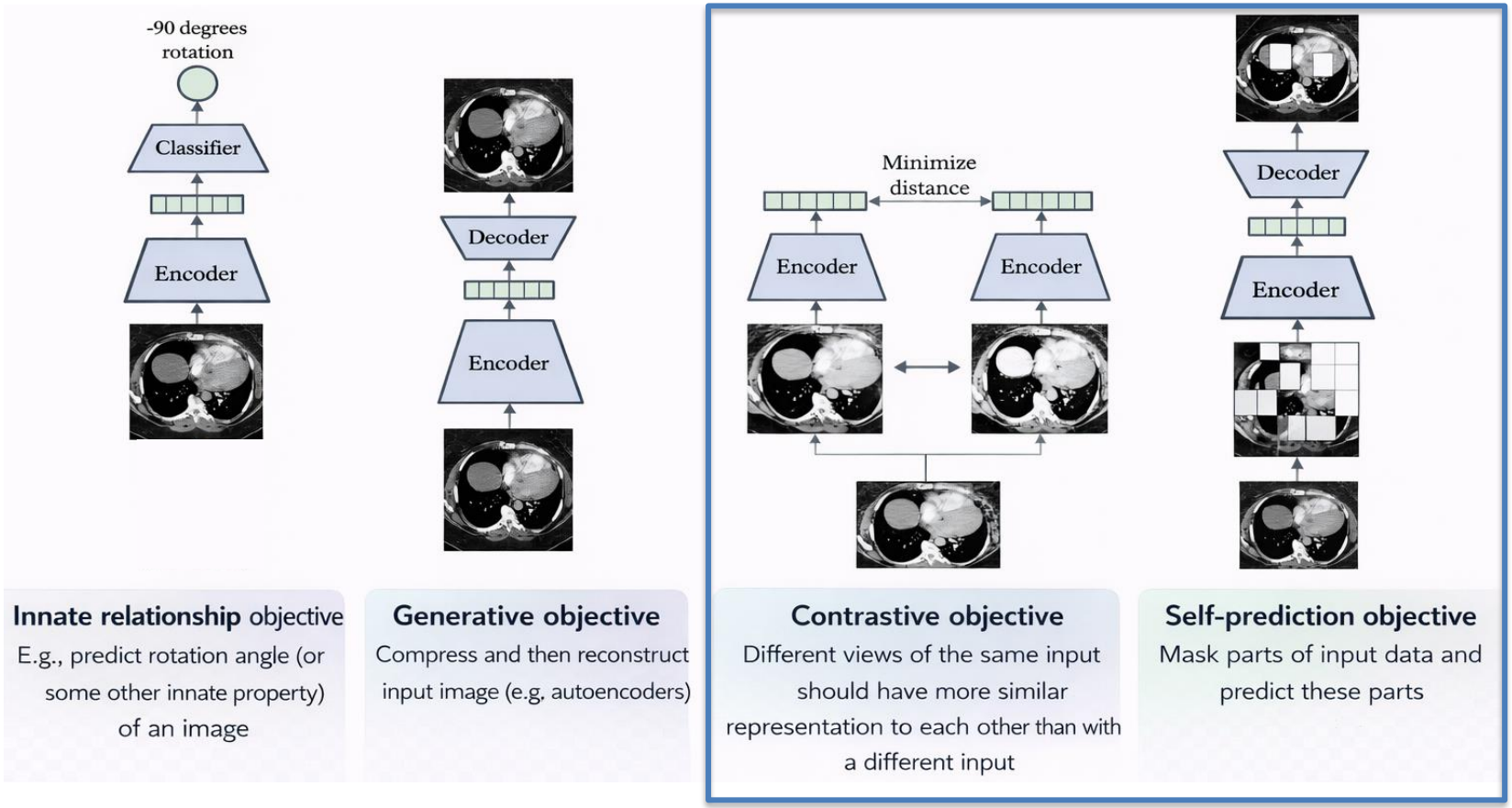
Generative objective
Compress and then reconstruct input image (e.g. autoencoders)

Contrastive objective
Different views of the same input should have more similar representation to each other than with a different input

Self-prediction objective
Mask parts of input data and predict these parts

Different representation learning paradigms

Popular State-of-the-art approaches



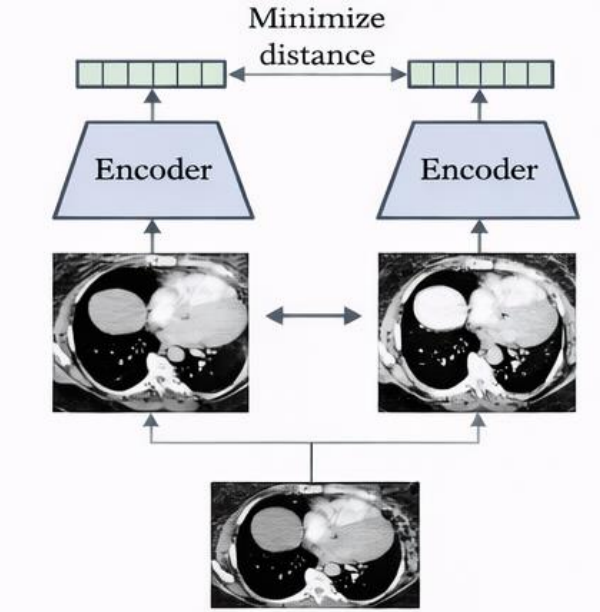
Innate relationship objective
E.g., predict rotation angle (or some other innate property) of an image

Generative objective
Compress and then reconstruct input image (e.g. autoencoders)

Contrastive objective
Different views of the same input should have more similar representation to each other than with a different input

Self-prediction objective
Mask parts of input data and predict these parts

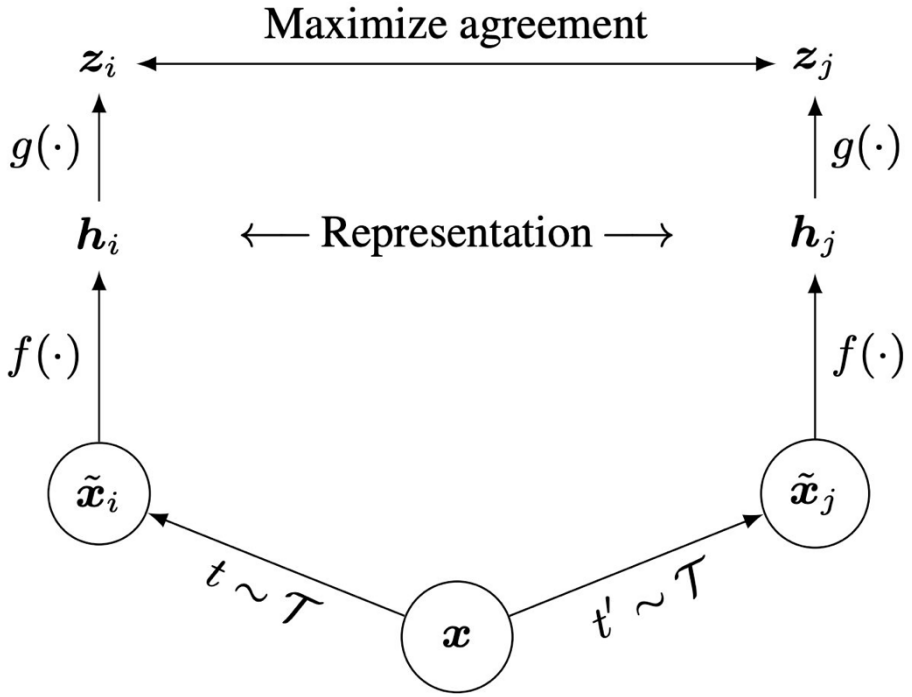
SimCLR: Contrastive Representation Learning



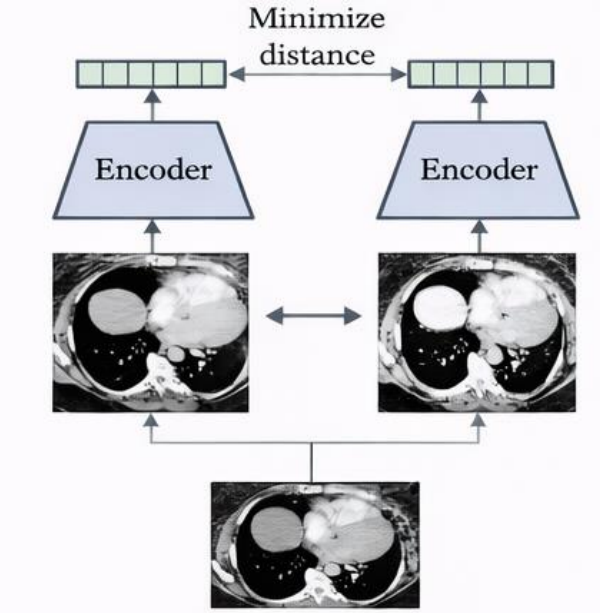
Contrastive objective

Different views of the same input should have more similar representation to each other than with a different input

SimCLR: "Simple Framework for Contrastive Learning of Visual Representations"



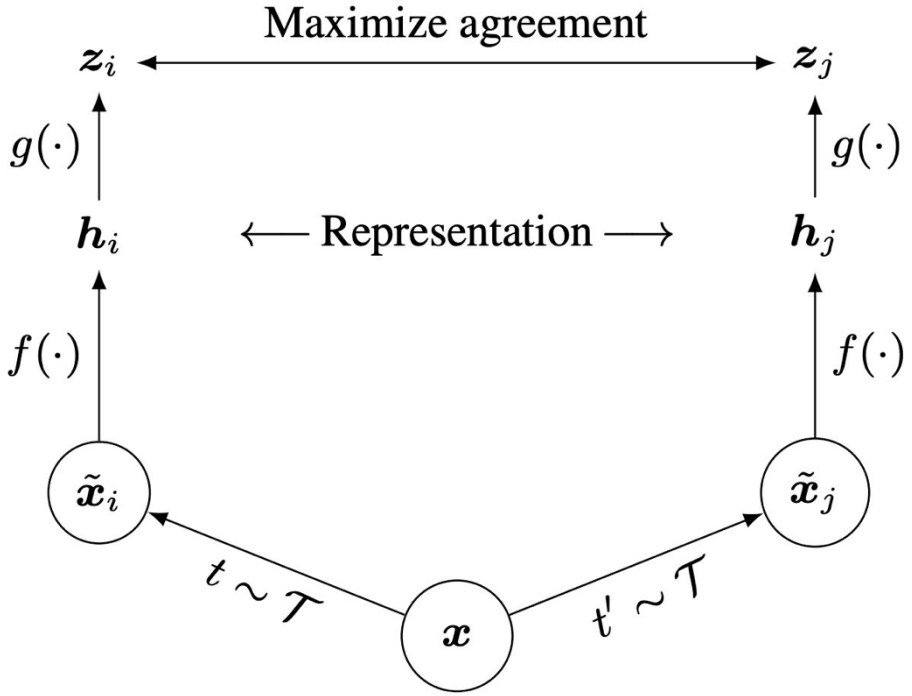
SimCLR: Contrastive Representation Learning



Contrastive objective

Different views of the same input should have more similar representation to each other than with a different input

SimCLR: "Simple Framework for Contrastive Learning of Visual Representations"



After self-supervised training, can fine-tune the encoder f on smaller labelled datasets. Can also directly extract learned representations h for downstream tasks.

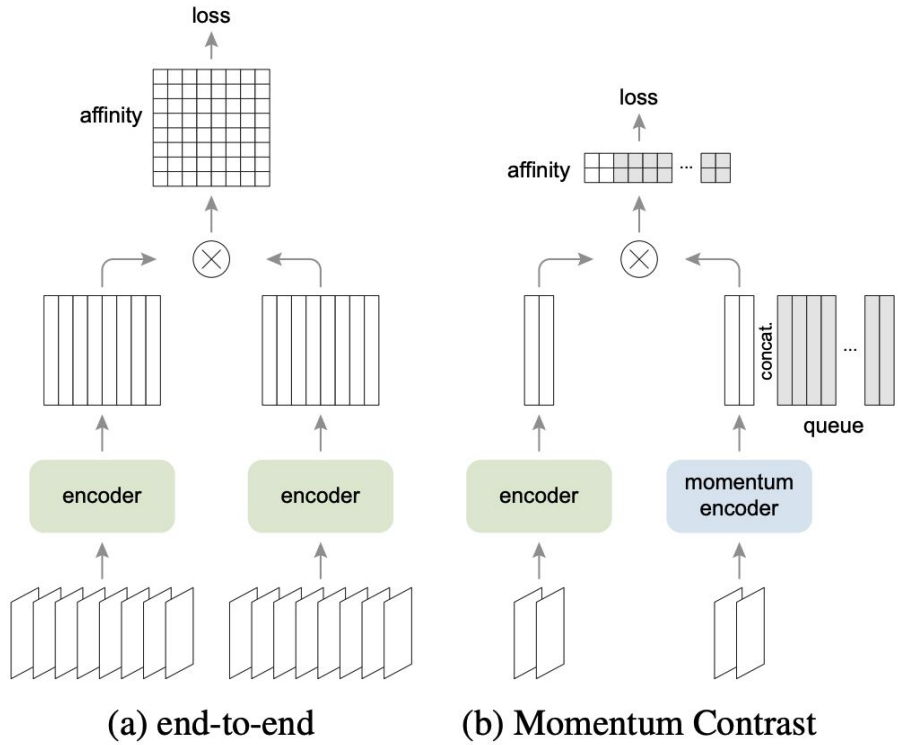
Limitations of SimCLR

- SimCLR's relies on a large batch size to generate a sufficient number of negative pairs for effective contrastive learning. This creates significant computational and memory burden.

MoCo: Alleviates the batch size limitation of SimCLR

- SimCLR's relies on a large batch size to generate a sufficient number of negative pairs for effective contrastive learning. This creates significant computational and memory burden.

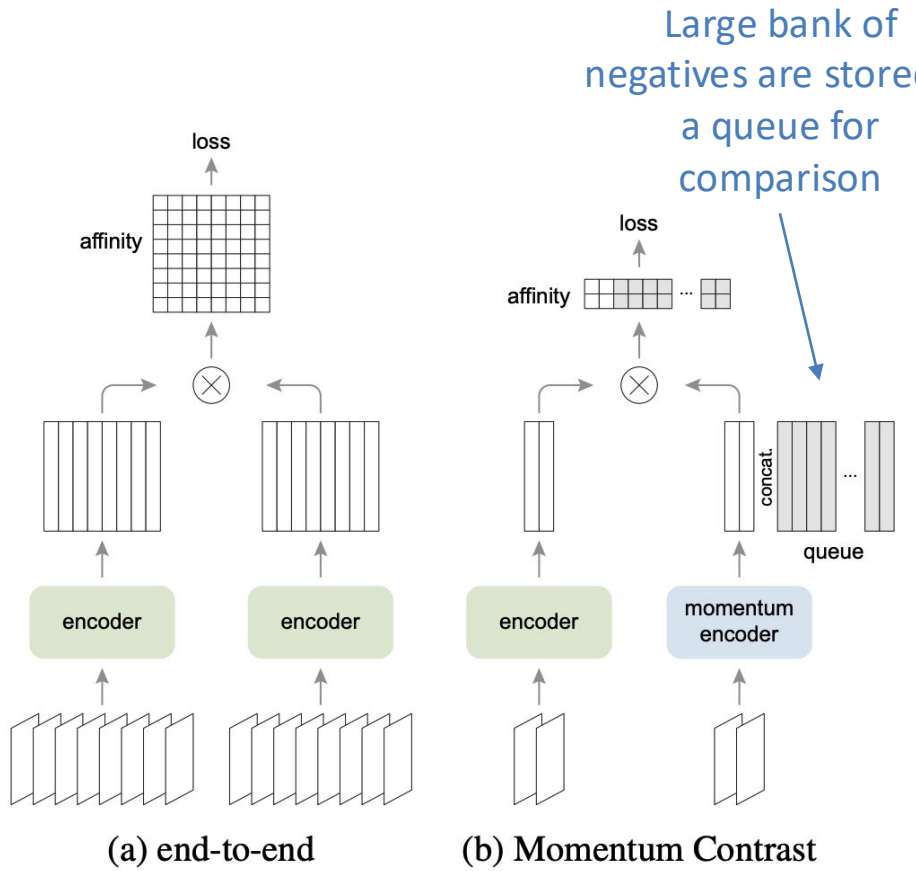
- MoCo (Momentum Contrast) and MoCo v2, v3 alleviates this by using a momentum-updated queue that allows incorporating many negative pairs without increasing batch size.



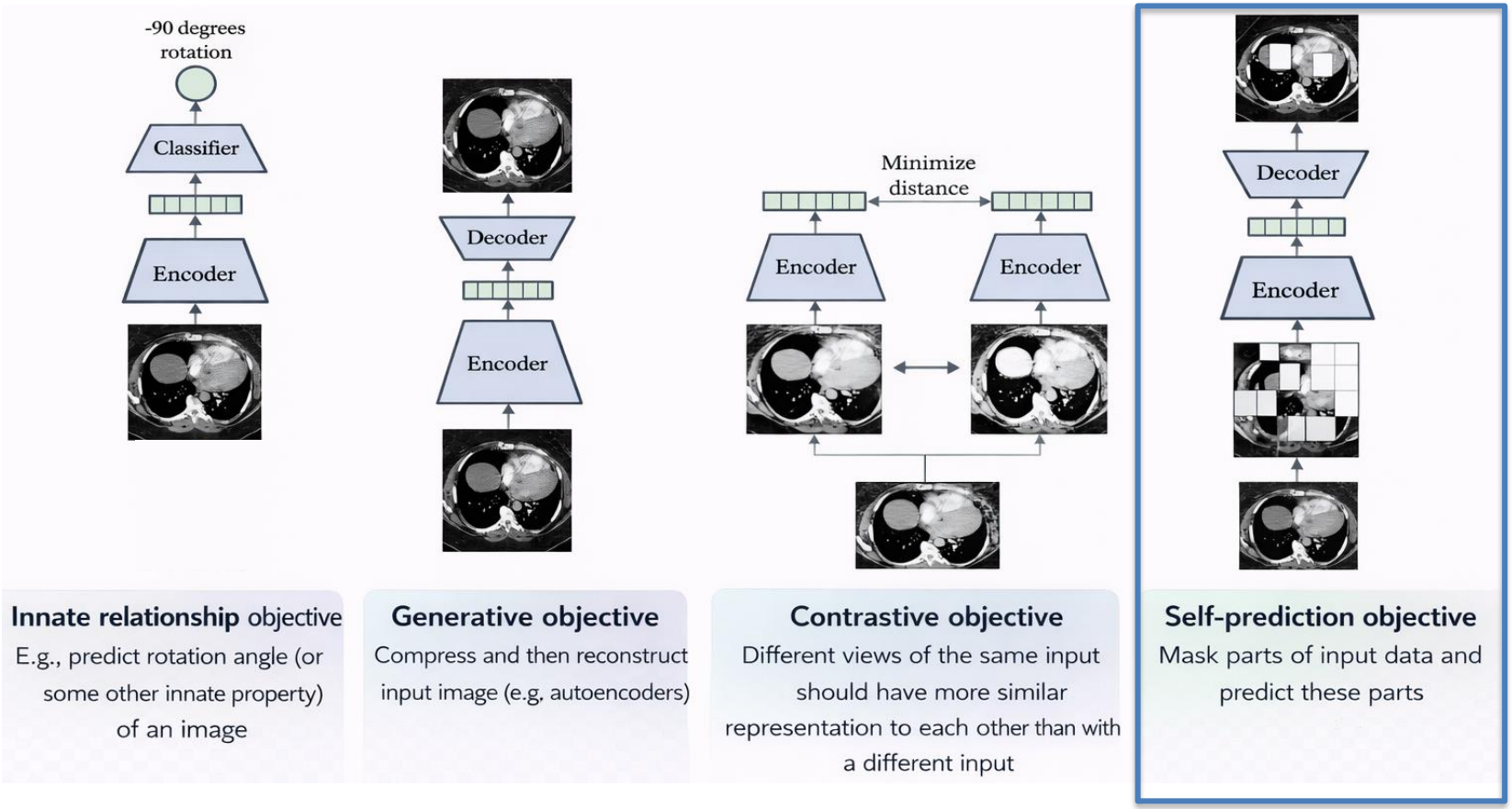
MoCo: Alleviates the batch size limitation of SimCLR

- SimCLR's relies on a large batch size to generate a sufficient number of negative pairs for effective contrastive learning. This creates significant computational and memory burden.

- MoCo (Momentum Contrast) and MoCo v2, v3 alleviates this by using a momentum-updated queue that allows incorporating many negative pairs without increasing batch size.



Different representation learning paradigms



Innate relationship objective
E.g., predict rotation angle (or some other innate property) of an image

Generative objective
Compress and then reconstruct input image (e.g. autoencoders)

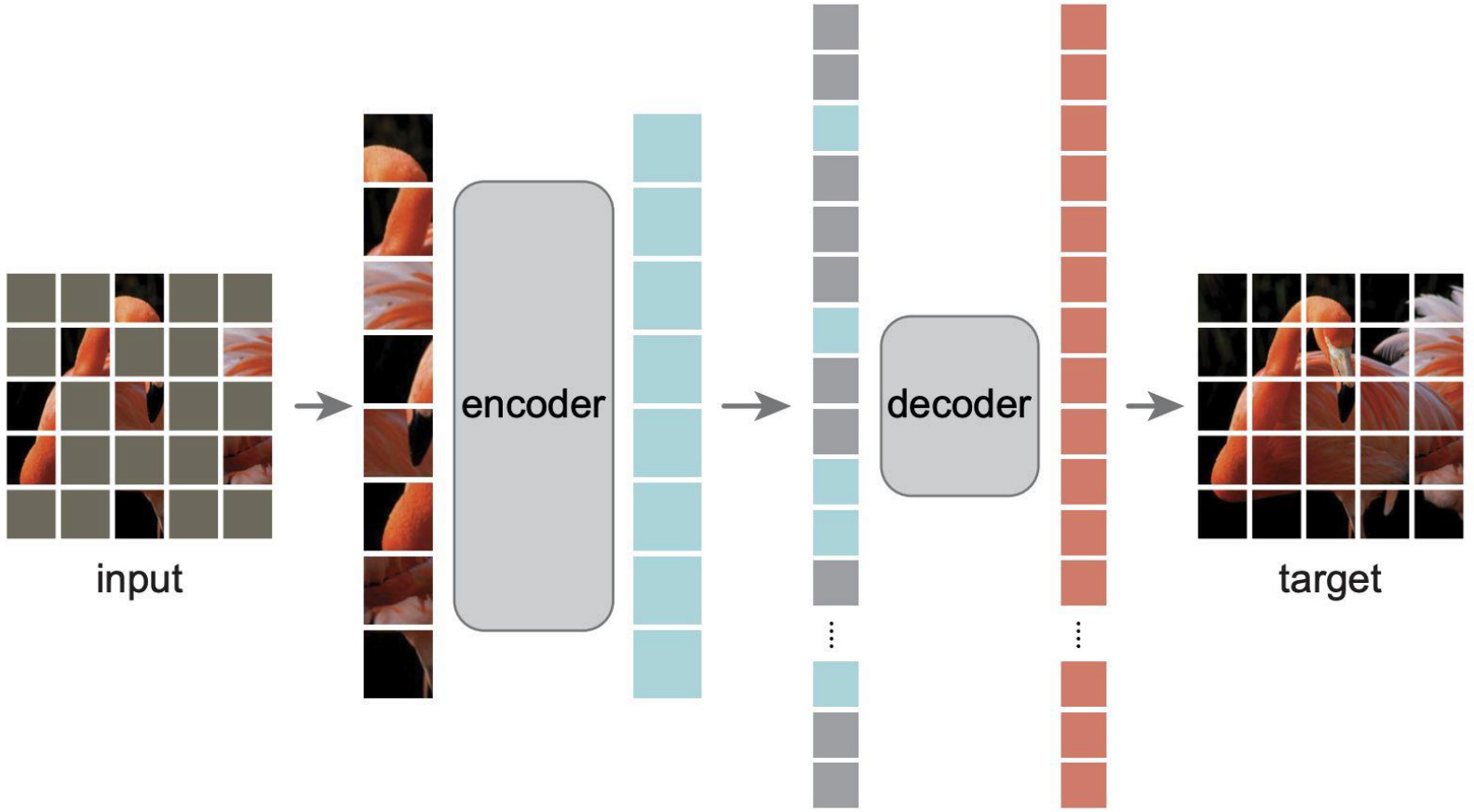
Contrastive objective
Different views of the same input should have more similar representation to each other than with a different input

Self-prediction objective
Mask parts of input data and predict these parts

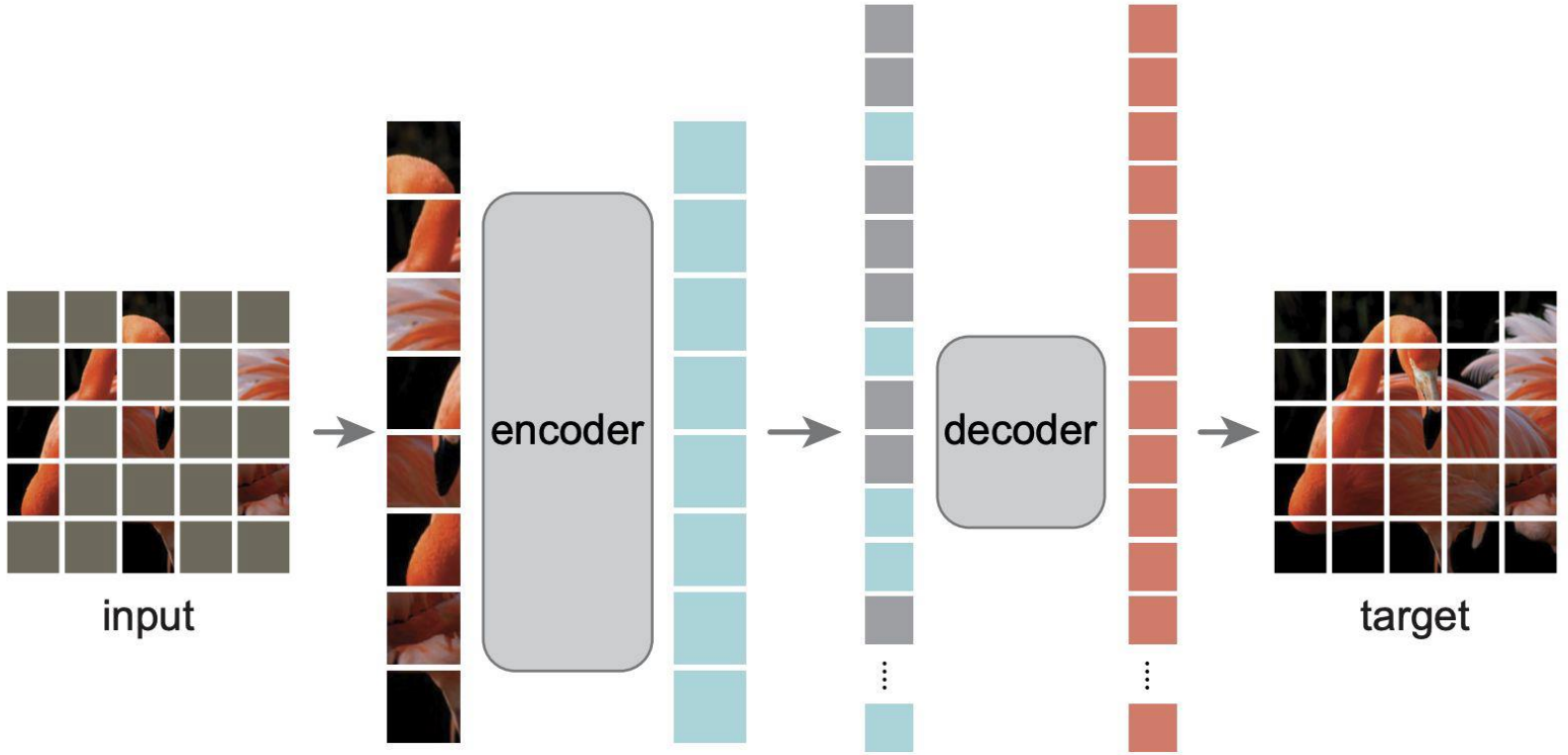
Masked Autoencoders (MAE)

- Key idea: instead of comparing images, we hide parts of the image and predict them
- Randomly mask image patches
- Train model to reconstruct missing content
- Learn semantic representations
- Inspired by major self-supervised representation learning paradigm in NLP (e.g. BERT), that masks tokens in sentences and trains models to reconstruct them
- Intuition: Transformer architecture is well-suited to this objective

Masked Autoencoders (MAE)

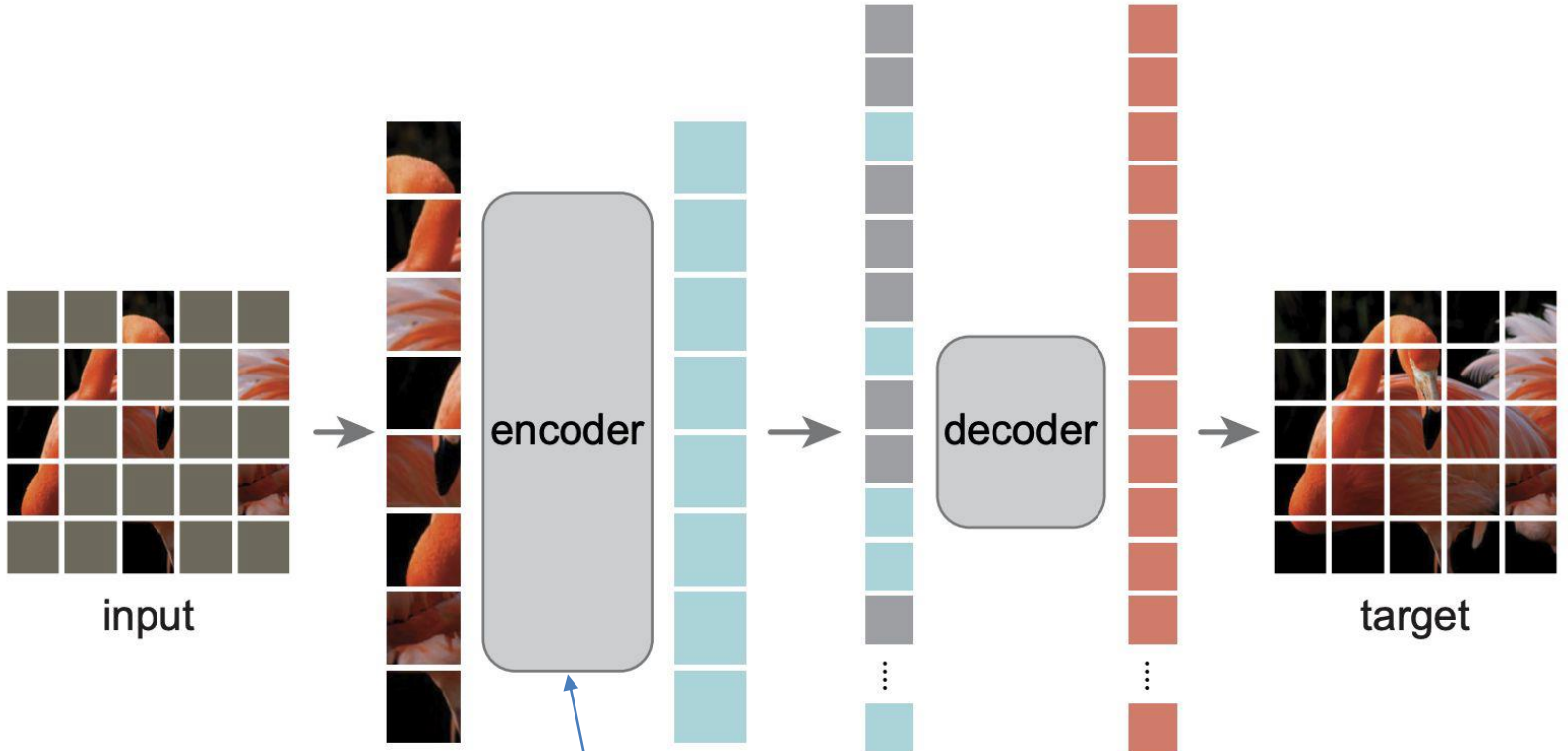


Masked Autoencoders (MAE)



Mask a large fraction of input image patches (e.g. 75%+)

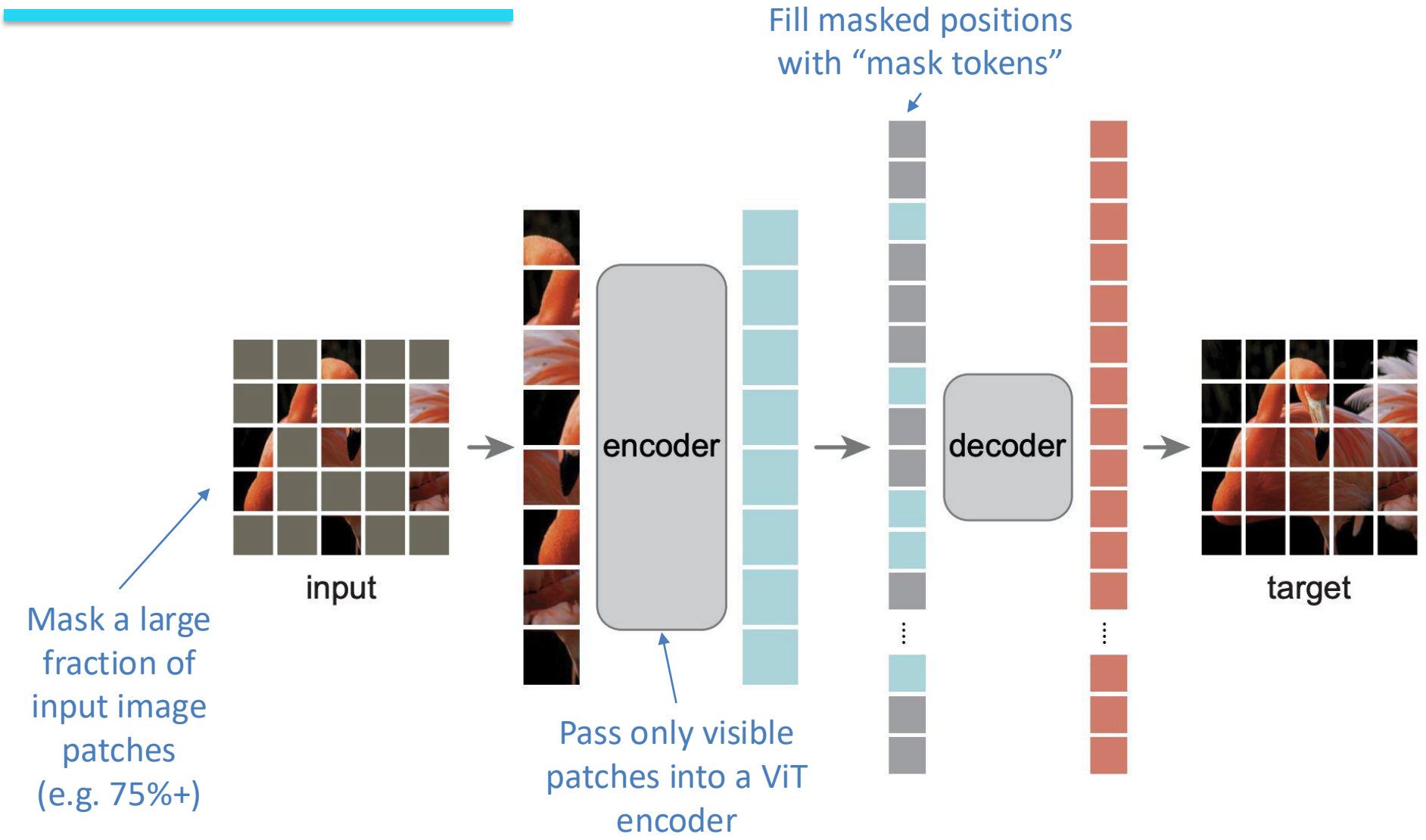
Masked Autoencoders (MAE)



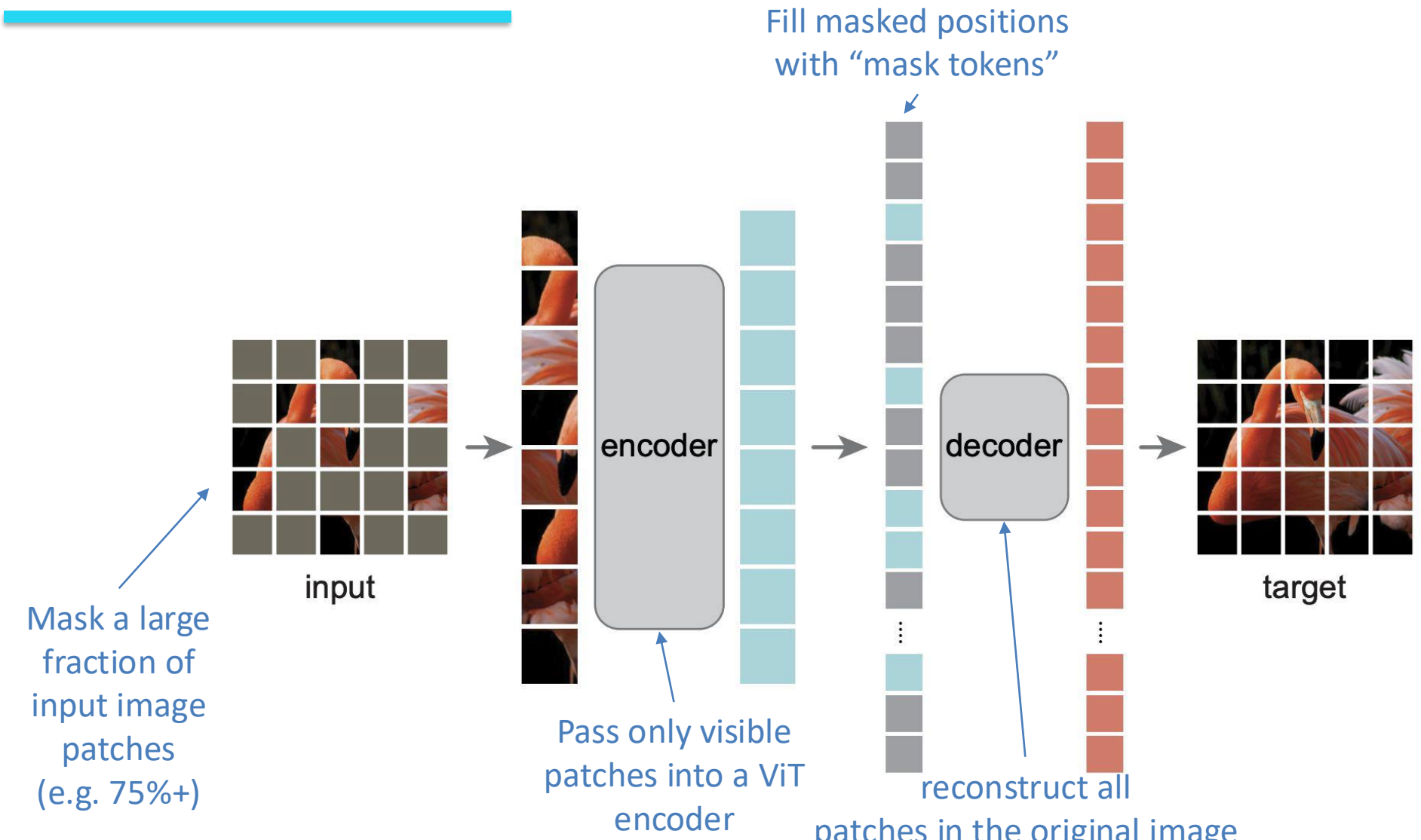
Mask a large fraction of input image patches (e.g. 75%+)

Pass only visible patches into a ViT encoder

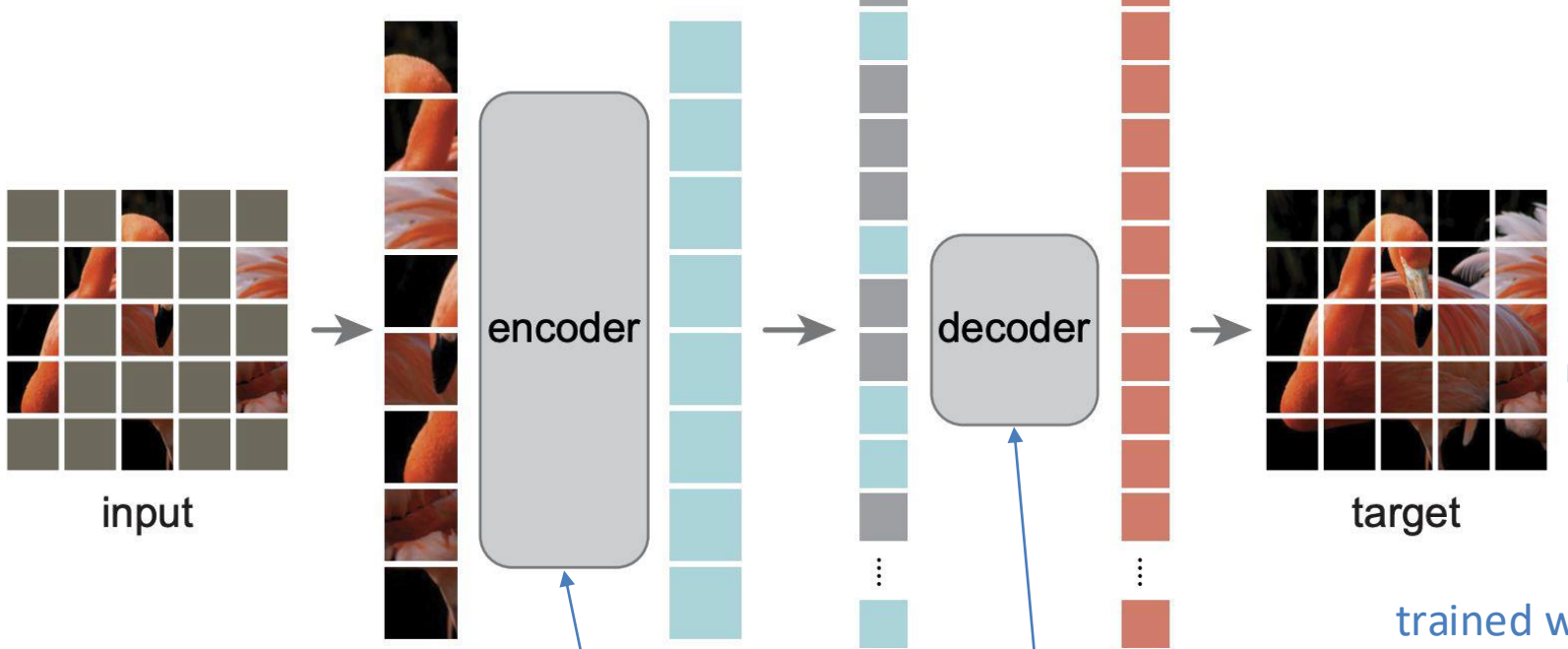
Masked Autoencoders (MAE)



Masked Autoencoders (MAE)



Masked Autoencoders (MAE)



Fill masked positions with “mask tokens”

Mask a large fraction of input image patches (e.g. 75%+)

Pass only visible patches into a ViT encoder

reconstruct all patches in the original image

trained with a reconstruction loss, MSE (mean squared error) between output and target image

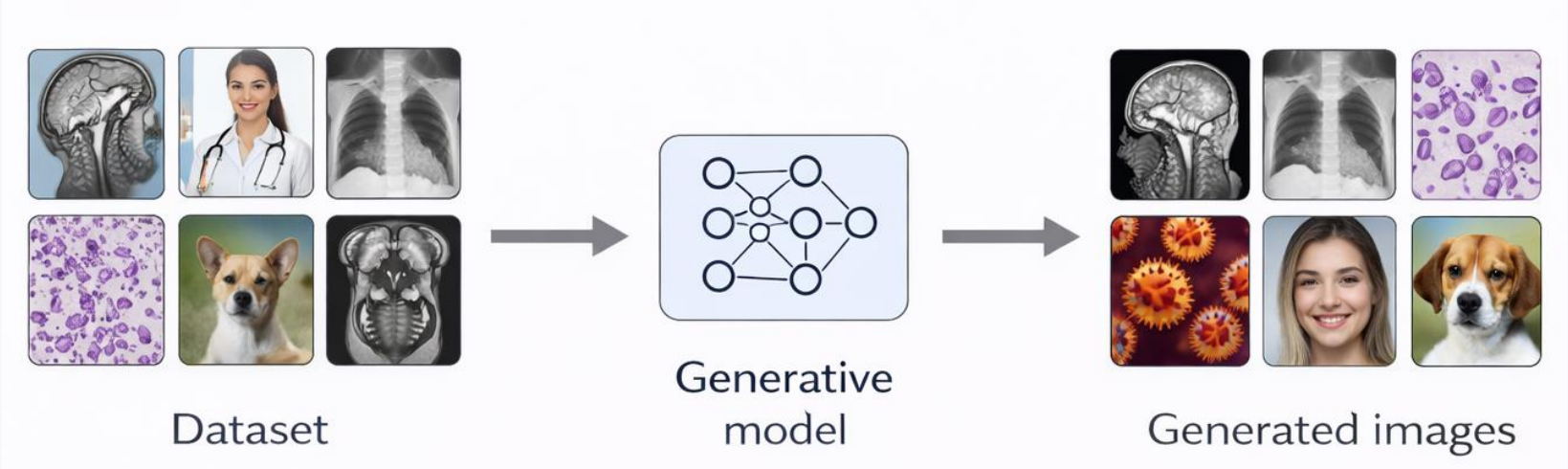
Self-Supervised Representation Learning

Method	Key Idea
SimCLR	Contrast augmented views
MoCo	Queue of negative examples
MAE	Mask and reconstruct image

- Learn representations from **unlabelled images**
- Different training strategies
- Representations transfer to **downstream tasks**

Part IV — Generative Models

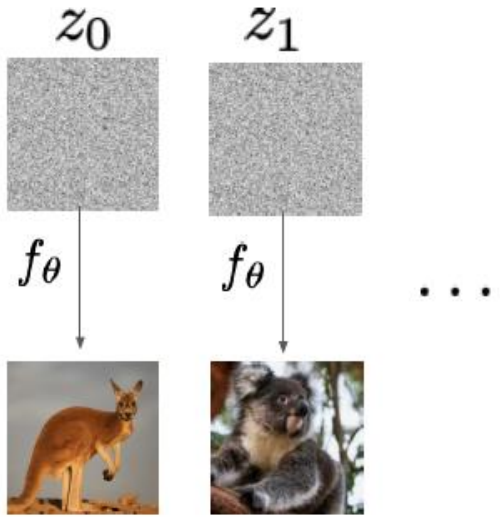
Vision Generative Models



- Representation learning focuses on extracting meaningful features from images
- Generative models aim to learn the data distribution
- This allows the model to generate new synthetic images

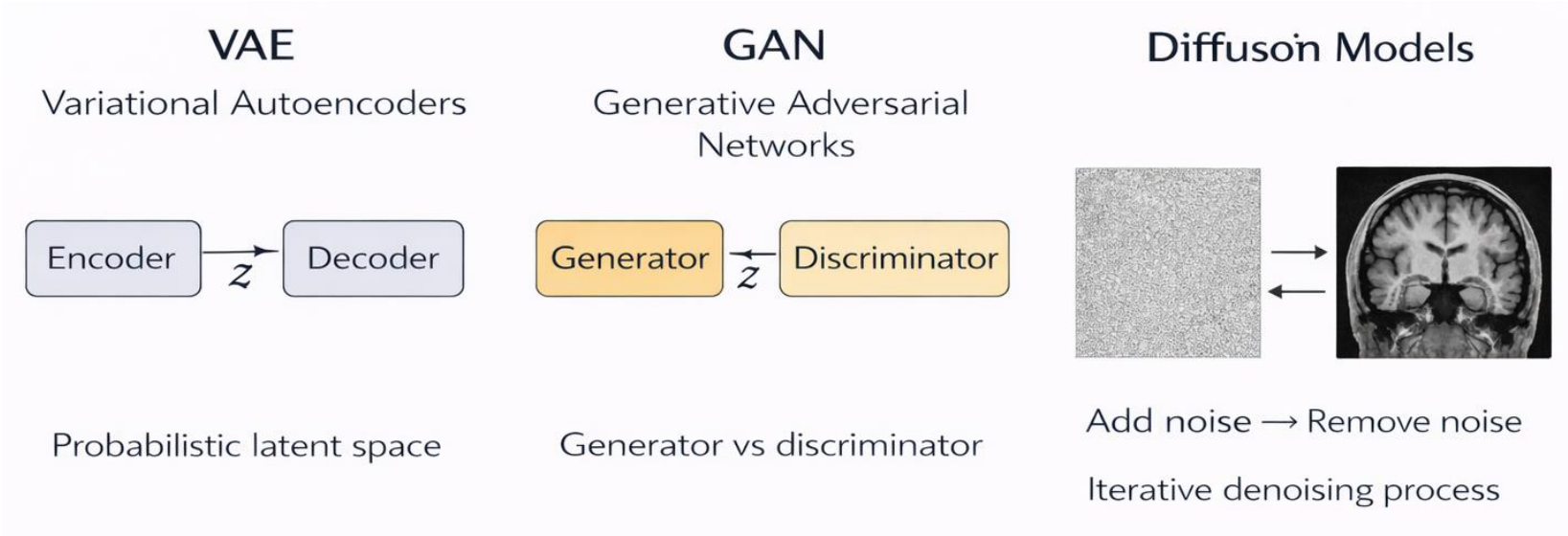
Generative modelling: the goal

Training



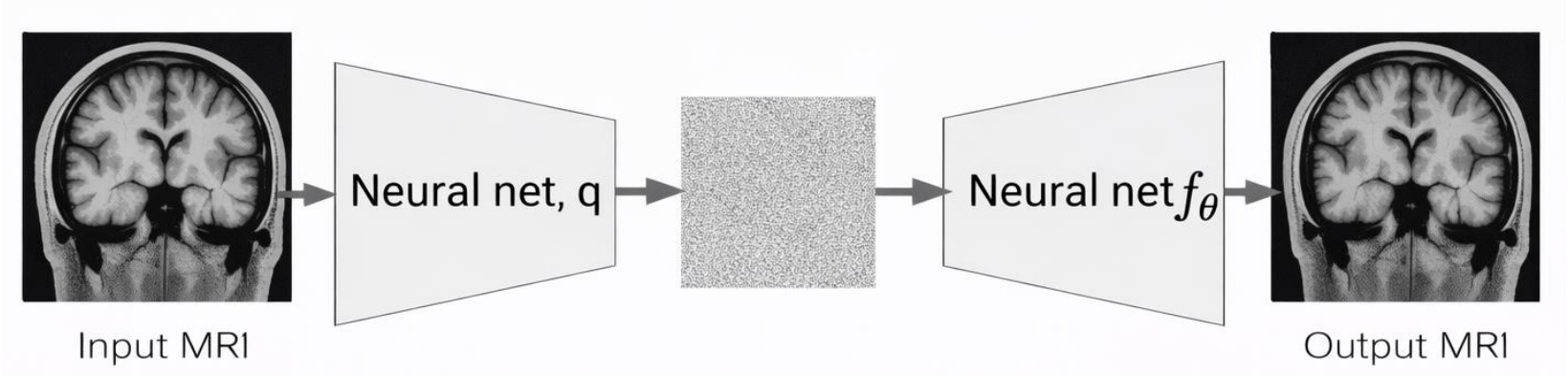
Generation

Families of Generative Models



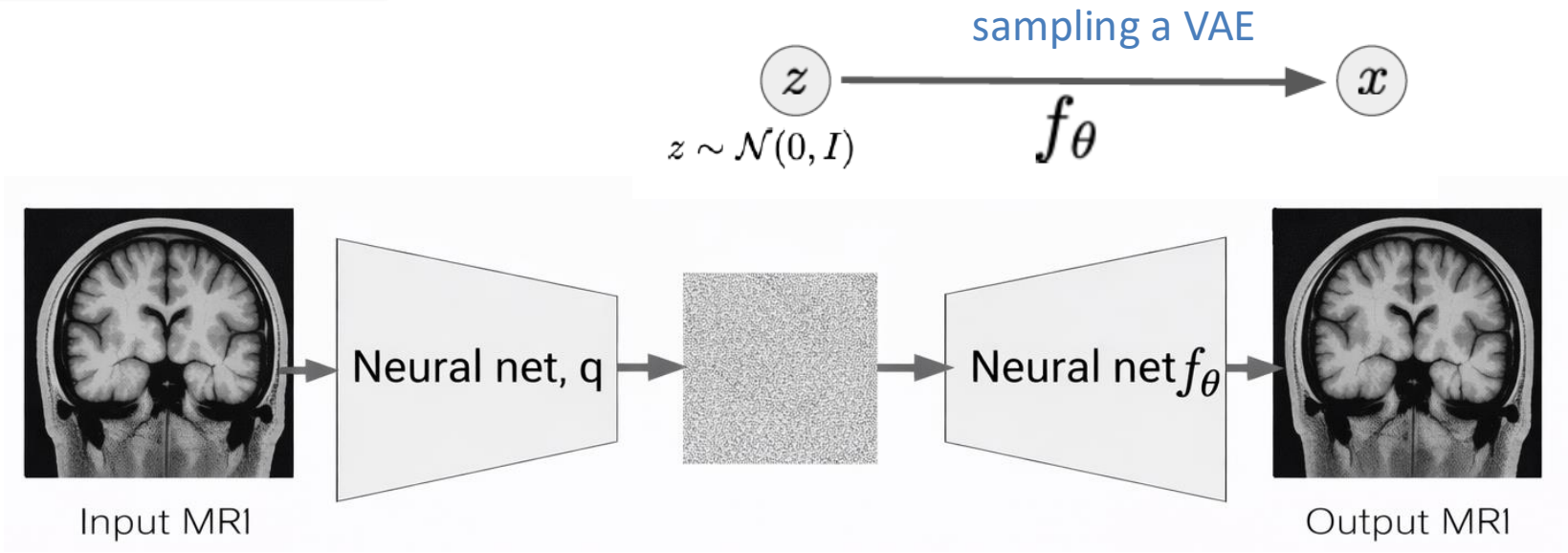
- Multiple approaches exist for modelling **data distributions**
- Each model uses a **different training strategy**
- We will look at three widely used methods

Variational Autoencoder (VAE)



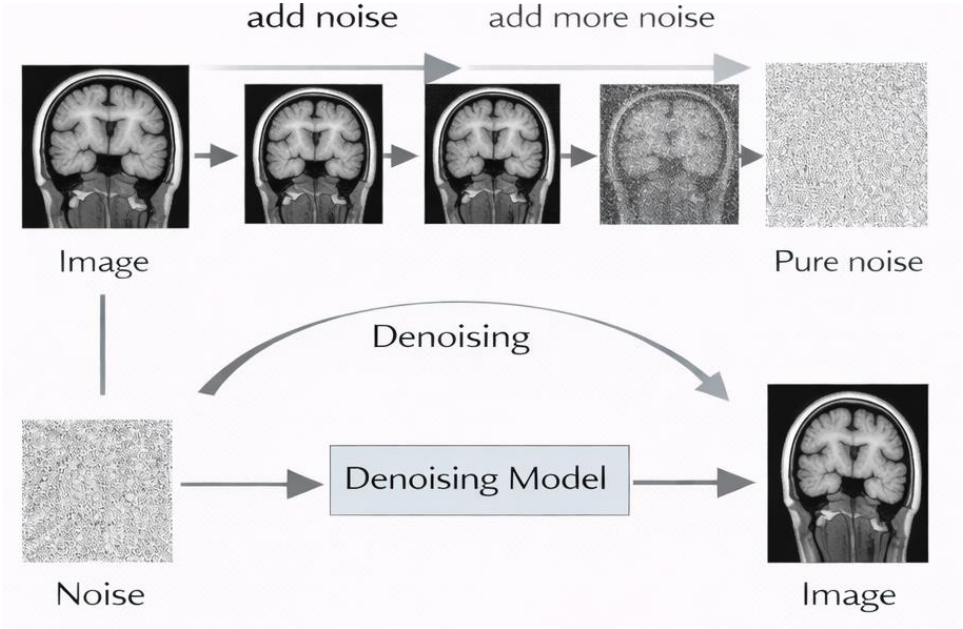
- Learn a probabilistic latent representation
- Encode images into a latent distribution
- Generate new samples by sampling latent vectors

Variational Autoencoder (VAE)



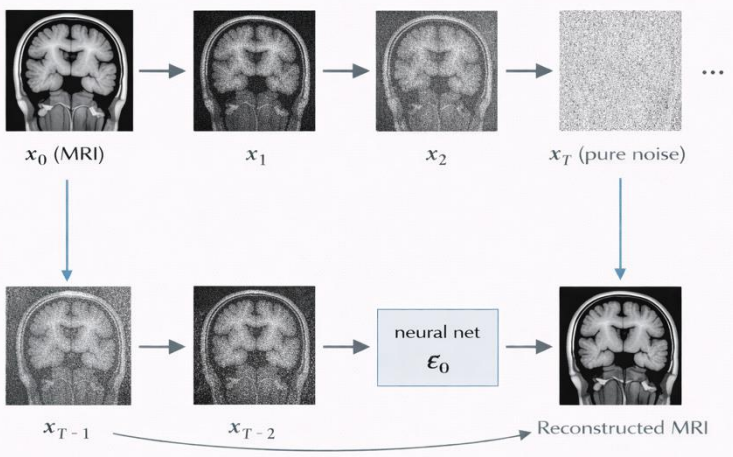
- Learn a probabilistic latent representation
- Encode images into a latent distribution
- Generate new samples by sampling latent vectors

Diffusion Models



- Forward process gradually **adds noise**
- Reverse process **removes noise step by step**
- Final output is a realistic synthetic image
- In medical imaging, diffusion models can generate synthetic MRIs, improve image quality, or reconstruct missing data

Learning to Predict Noise

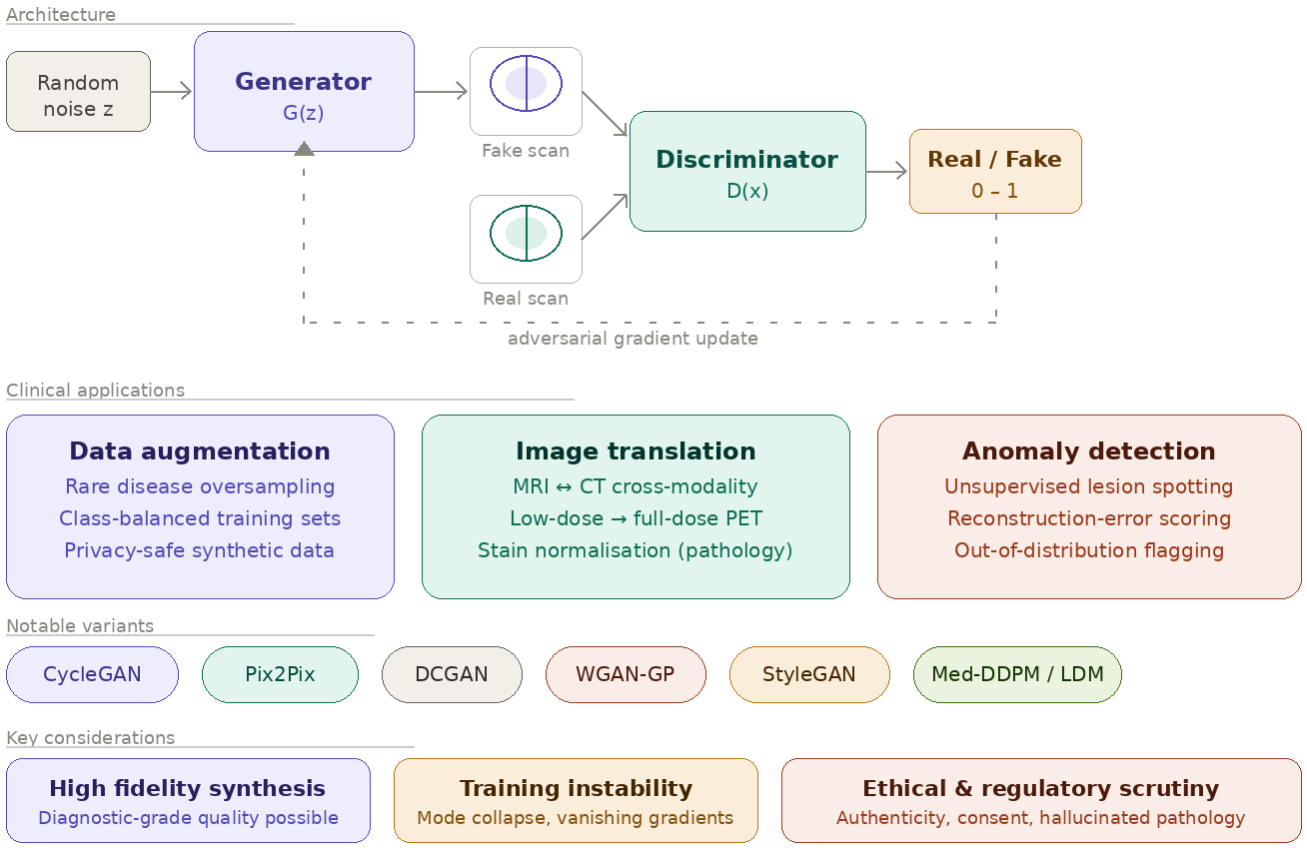


- During **training**, we start with a real image. We gradually add Gaussian noise over many steps
- As we add noise step by step, the image becomes increasingly corrupted
- After many steps, the image becomes pure noise
- The key idea of diffusion models is to learn the reverse process
- Starting from noisy inputs, the model learns how to gradually remove noise
- A neural network is trained to predict the noise present in the image at each step
- During **generation**, we start from pure noise and repeatedly apply the learned denoising process until a realistic image is produced

Diffusion Models in Medical Imaging

- **Image reconstruction**: recover high-quality images from undersampled or noisy acquisitions (for example, accelerated MRI or low-dose CT)
- **Denoising**: remove acquisition noise while preserving clinically relevant anatomical structure
- **Modality translation**: generate one modality from another, such as MRI → CT
- **Synthetic data generation**: create realistic medical images for training, augmentation, and rare-case simulation
- Important diffusion model designs
  - Denoising Diffusion Probabilistic Models (DDPM)
  - Latent Diffusion Models (much more efficient)

Generative Adversarial Networks (GANs) in Medical Imaging



Part V — Emerging Directions & Discussion

Emerging Trends in Representation Learning

- Foundation models for medical imaging: Large pretrained vision models adapted to biomedical tasks
- Self-supervised learning at scale: Learning representations from large unlabelled medical datasets
- Multimodal representations: Combining imaging with clinical data, genomics, or text
- Vision-language models for medicine: Joint learning from images and radiology reports

Emerging Trends in Generative Models

- Diffusion models for high-quality image synthesis
- Controllable generation: Generating images with specific clinical attributes
- Synthetic data for training AI models
- Generative models for image reconstruction and enhancement

Key Takeaways

- Representation learning extracts useful features from images
- Self-supervised learning enables training with limited labels
- Generative models can simulate and reconstruct biomedical images
- These techniques are becoming central tools in AI for healthcare

Questions for Discussion

Where is the first real impact?

What is the biggest bottleneck?

Which direction matters most next?

How do we build trust?

From **impressive models** to real biomedical value.

Key Insights

- Diagnostic imaging is the best first impact area.
- Data quality and evaluation remain major bottlenecks.
- Multimodal foundation models will be game changers.
- Trust requires robustness, interpretability, and validation.