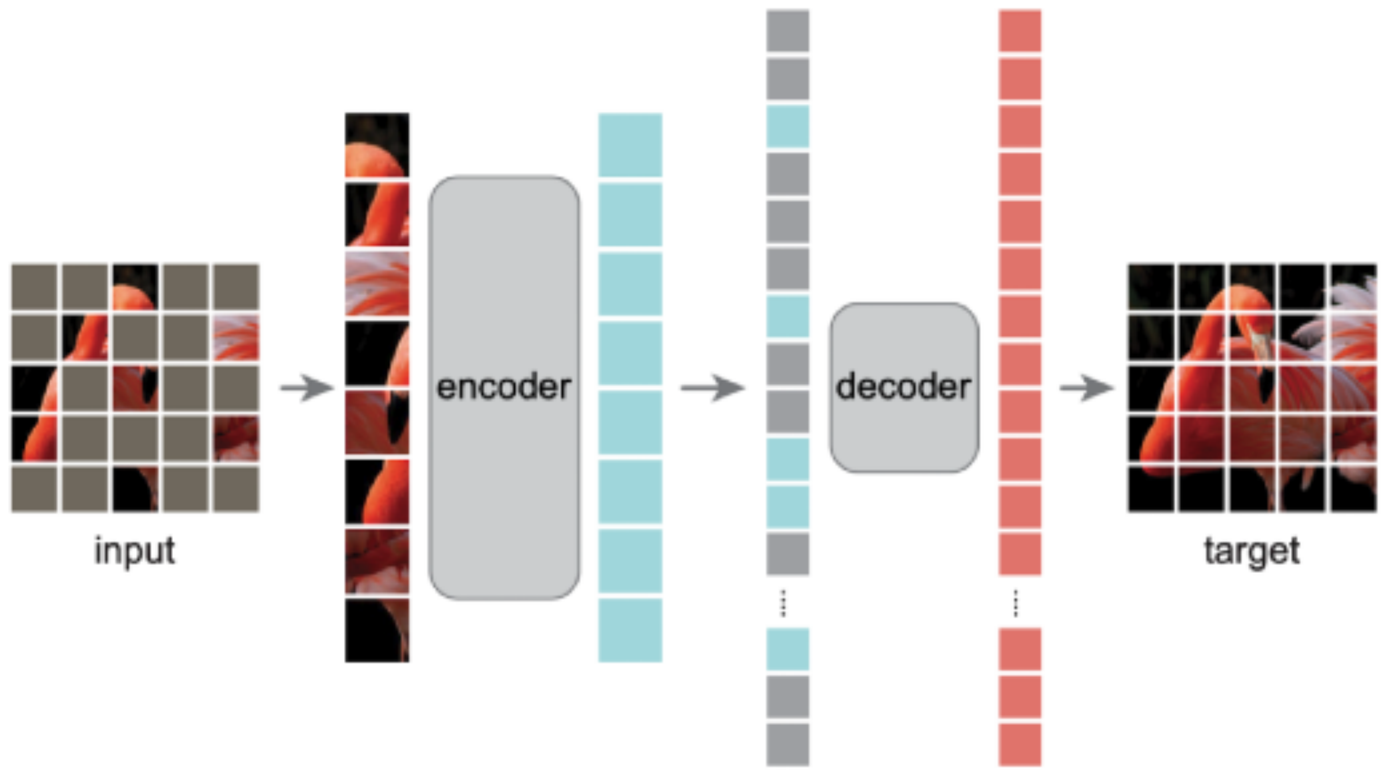
Masked Auto-encoders

Learning with fill in the blanks

The "How":

- *Mask:* A large portion of the input (e.g., 75% of patches)
- *Encode:* A deep Encoder processes *only the visible patches*.
- Reconstruct: A Decoder guess the missing patches.



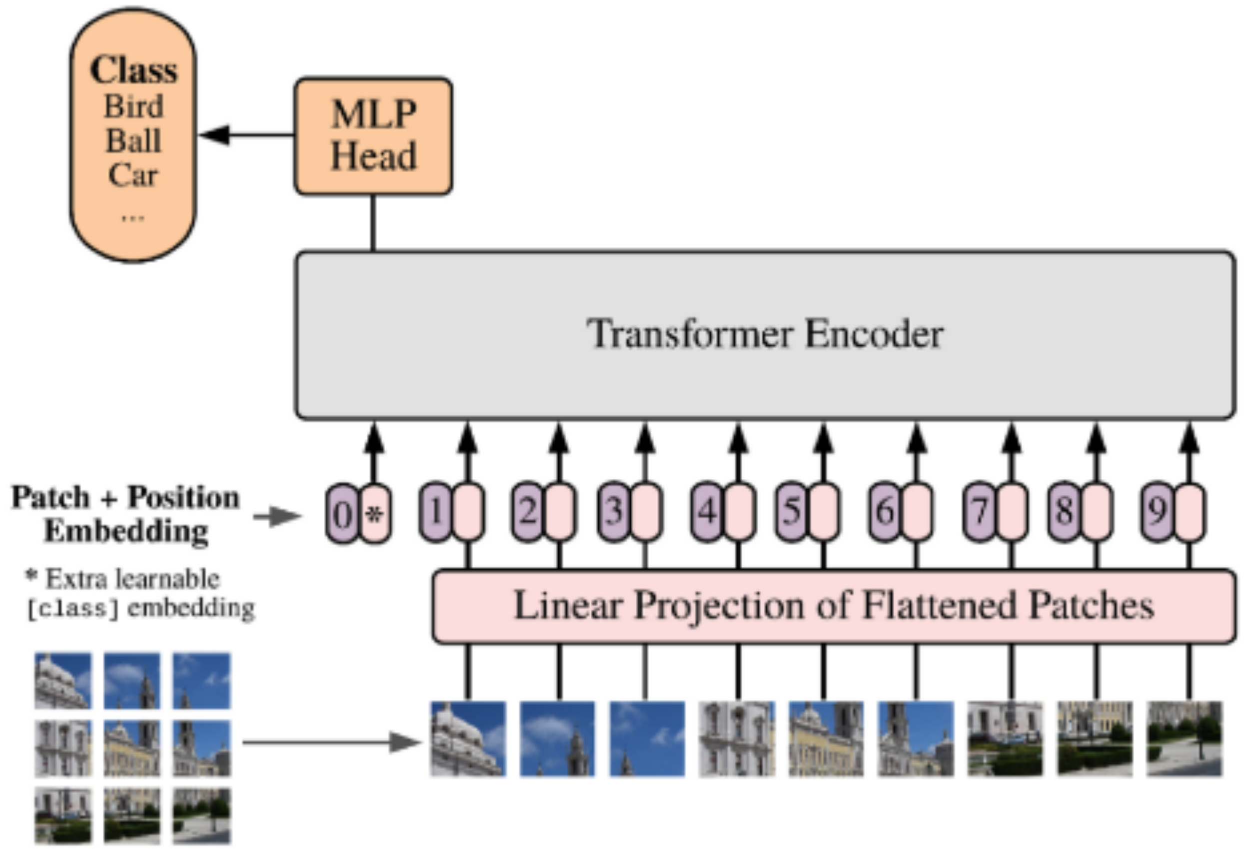
- **The Goal:** force Encoder to learn a rich representation of the data, not just surface-level details.

Vision Transformer

Self-attention to see the "big picture"

The "How":

- *Patchify*: An image is broken down into a sequence of patches.
- *Embed*: Each patch is converted to feature vector + positional info
- *Transformer Encoder*: self-attention to model the token relation



- **The Goal:** capture long-range dependencies and global context across the entire input.

Our Training Strategy

A Two-Stage Approach

- **Stage 1: Self-Supervised Pre-Training**
  - *Goal*: Force the model to learn a rich, physical representation of events.
  - *How*: A dual-objective Masked Autoencoder (MAE).
  - *Reconstruction Task:* Reconstruct masked (hidden) parts of the event.
  - *Contrastive Task:* Group hits that belong to the same voxel ID.
- **Stage 2: Supervised Fine-Tuning**
  - Goal: Adapt the "smart" pre-trained encoder to specific physics tasks.
  - How: Use the pre-trained weights as a starting point and fine-tune on the labeled dataset for classification and regression.