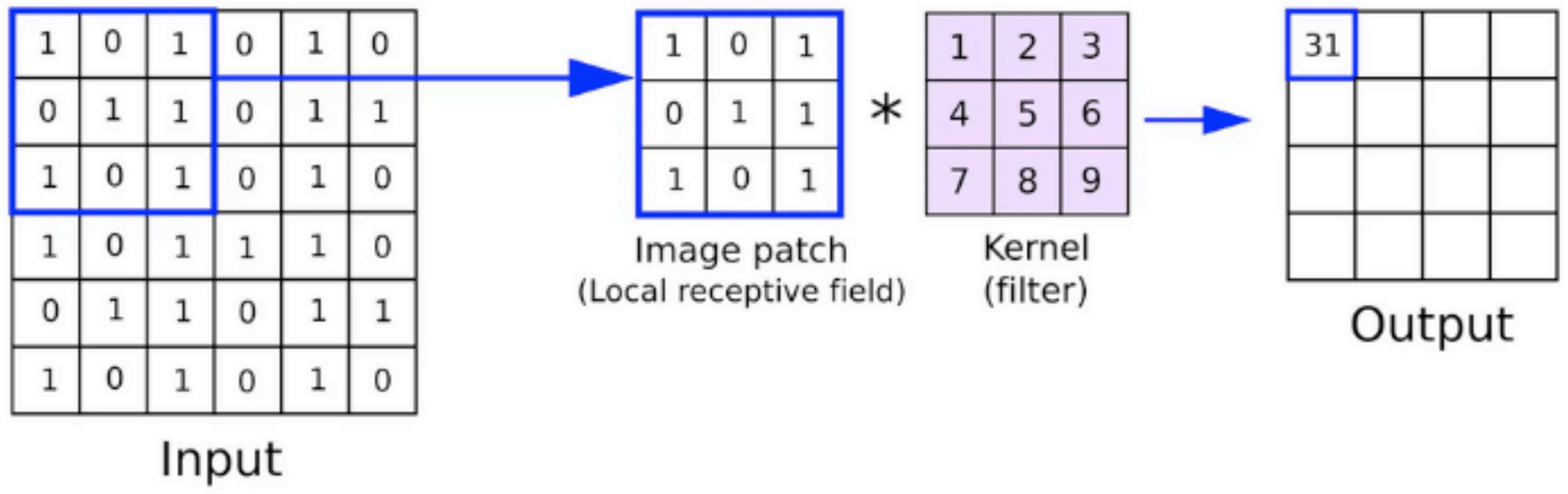
Cool, but How?

A Two-Stage Training Strategy

A Hybrid Model to Capture Local & Global Information:

- **SSCN (Local Info):**
  - A convolution that operates only on active voxels.
  - Efficiently learns local 3D features (shower shapes, track segments).
- **Hierarchical Transformer (Global Info):**
  - **Intra-Module Attention:** Summarizes patterns *within* each detector module.
  - **Inter-Module Attention:** Combines module summaries to learn the *entire event topology*

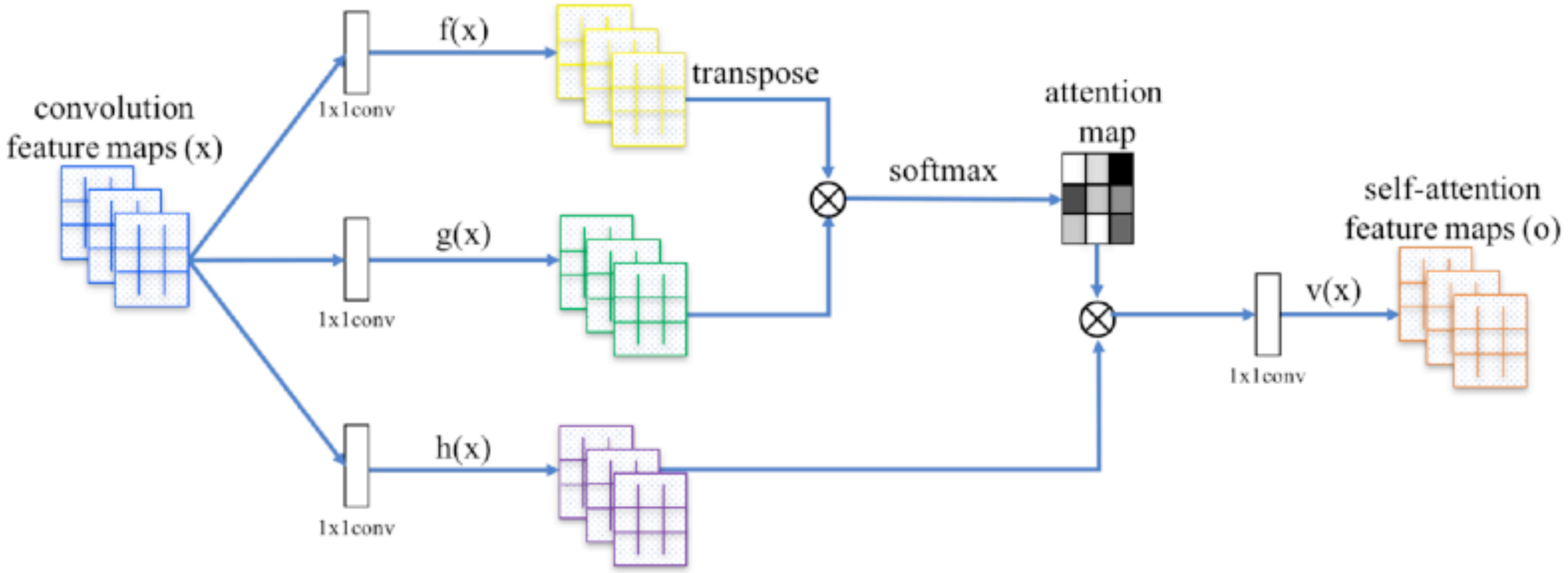


1	0	1	0	1	0
0	1	1	0	1	1
1	0	1	0	1	0
1	0	1	1	1	0
0	1	1	0	1	1
1	0	1	0	1	0

1	0	1
0	1	1
1	0	1

1	2	3
4	5	6
7	8	9

31			

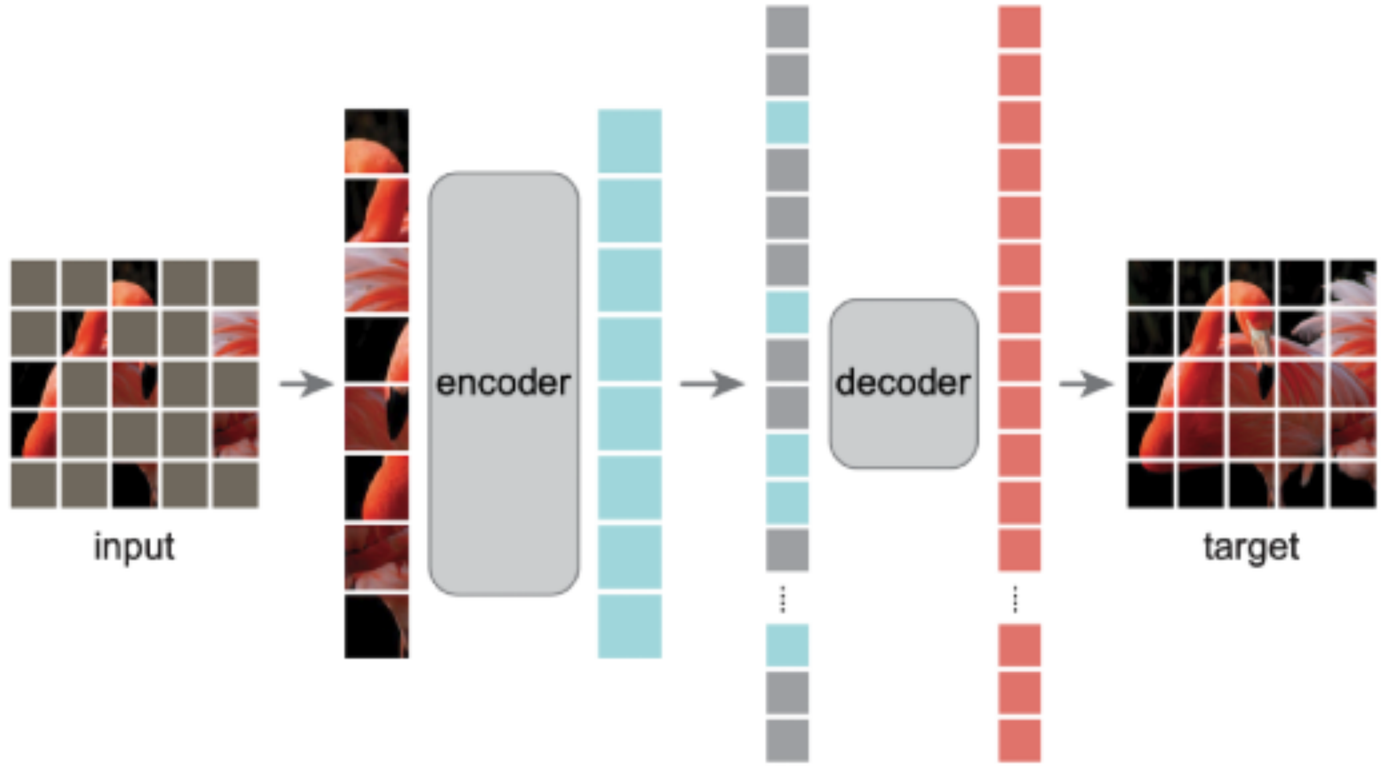


Masked Auto-encoders

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.



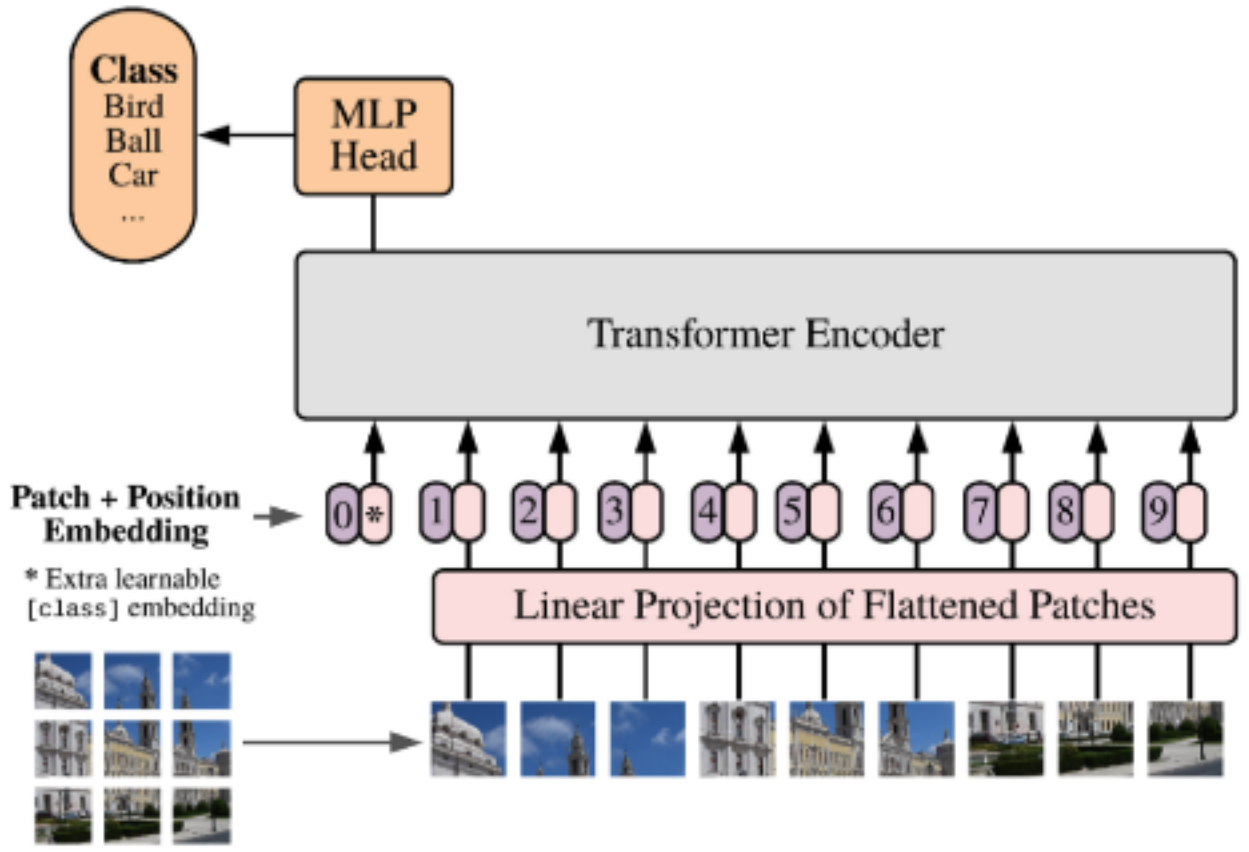
- **The Goal:** force Encoder to learn a rich representation of the data, not just surface-level details.

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


Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0 *
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

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