# Autoencoder

Masked Autoencoders are scaleable vision learners

<https://arxiv.org/pdf/2111.06377.pdf>

Discrete VAE (Text-to-Image)

<https://arxiv.org/pdf/2102.12092.pdf>

# Aligned Learning Objectives

Data2Vec

https://arxiv.org/pdf/2202.03555.pdf

Data2Vec V2

<https://arxiv.org/pdf/2212.07525.pdf>

Data2Vec V2 + Speech Generation

<https://ieeexplore.ieee.org/document/10095591>

Deep Aligned Representations (Similar to Data2Vec)

<https://arxiv.org/pdf/1706.00932.pdf>

# Multimodal

VLMo (Mixture-of-Experts)

<https://arxiv.org/pdf/2111.02358.pdf>

BEiT-V3 (MoE, V-QVT)

<https://arxiv.org/pdf/2208.10442.pdf>

BeiT-V2 (MoE, V-QVT)

<https://arxiv.org/pdf/2208.06366.pdf>

BEiT (Visual Tokenizer)

<https://openreview.net/pdf?id=p-BhZSz59o4>

FLAVA

<https://arxiv.org/pdf/2112.04482.pdf>

Learning Transferable Visual Models From Natural Language Supervision (Image+Text)

<https://arxiv.org/pdf/2103.00020.pdf>

# General

Iterative Attention

<https://arxiv.org/pdf/2103.03206.pdf>

Acoustic Representation

<https://arxiv.org/pdf/1912.01679.pdf>