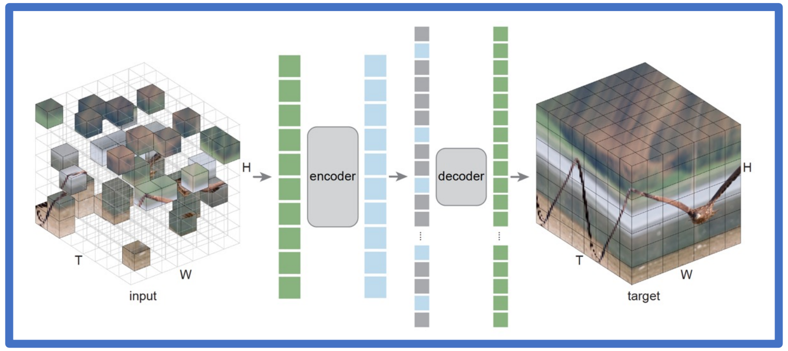
Self-supervised learning from masked video and audio

Christoph Feichtenhofer

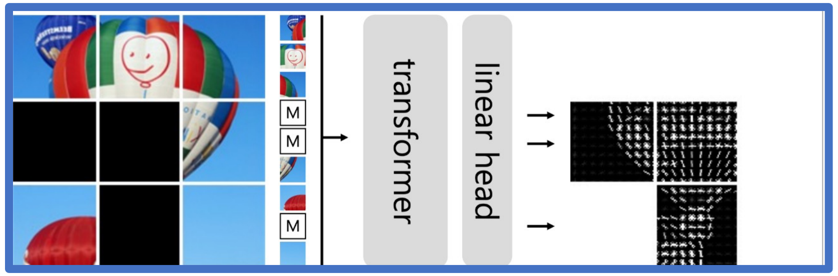
Meta AI, FAIR

Outline: Masked Video Representation Learning

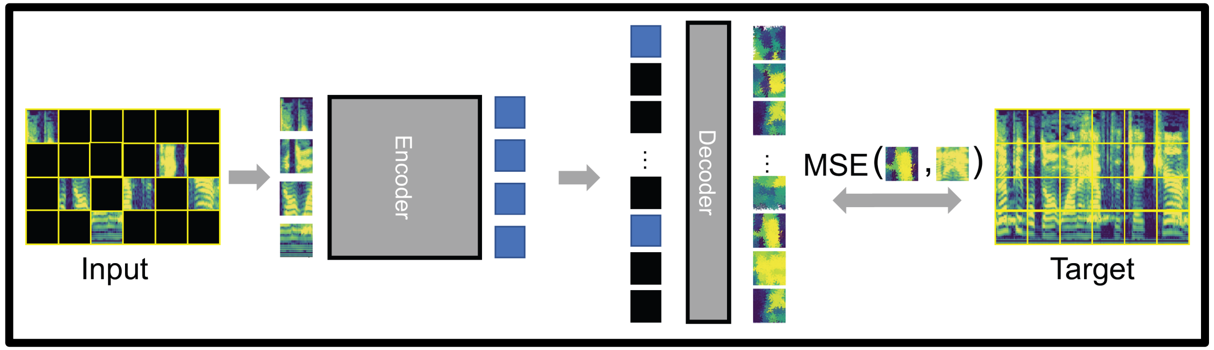
1. Masked Autoencoders (MAE) for video



2. MaskFeat studying features for masked autoencoding



3. Audio Learning with MAE



Masked Autoencoders As Spatiotemporal Learners

Christoph Feichtenhofer*, Haoqi Fan*, Yanghao Li, Kaiming He

Meta AI, FAIR

github.com/facebookresearch/mae_st
github.com/facebookresearch/SlowFast

Masked Language Modeling

men basketball

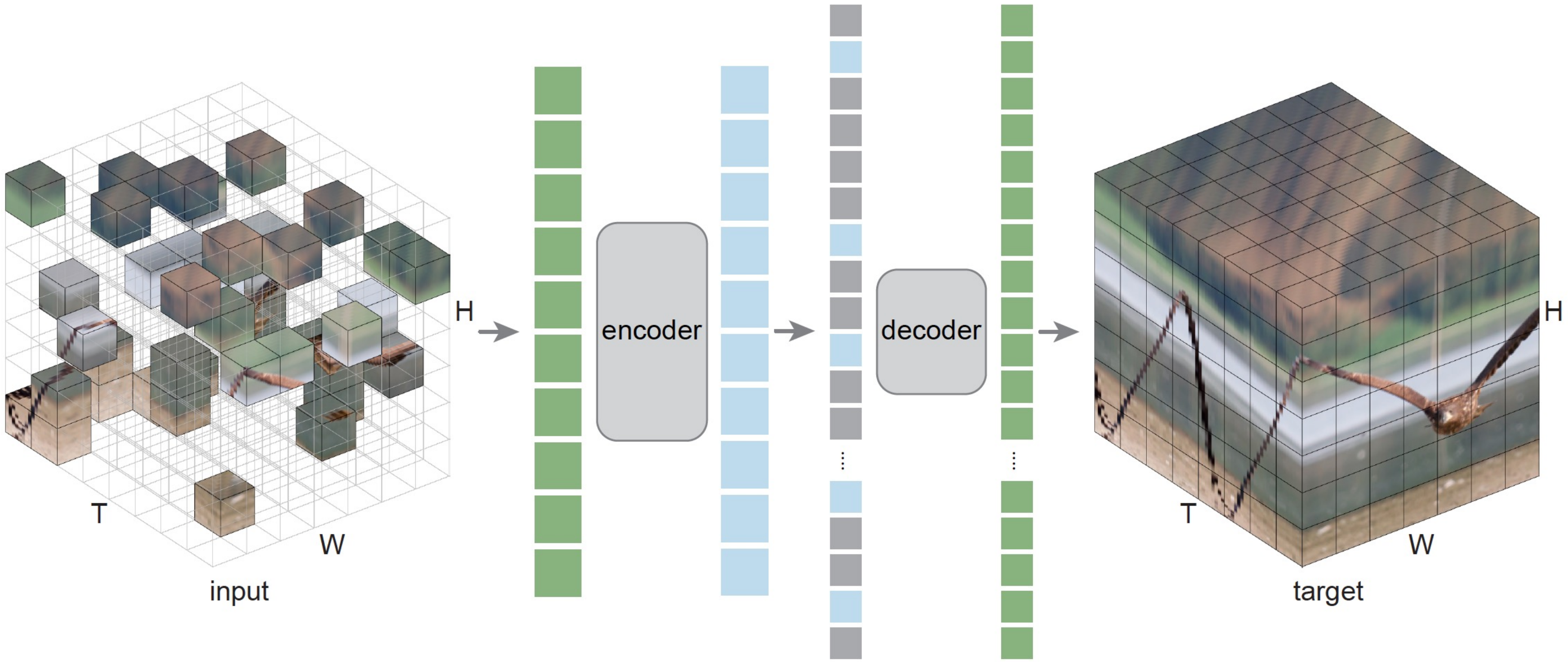
“Two playing a game of on an outside court.”

Devlin *et al.*, BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding

Masked Autoencoders (MAE) for visual learning

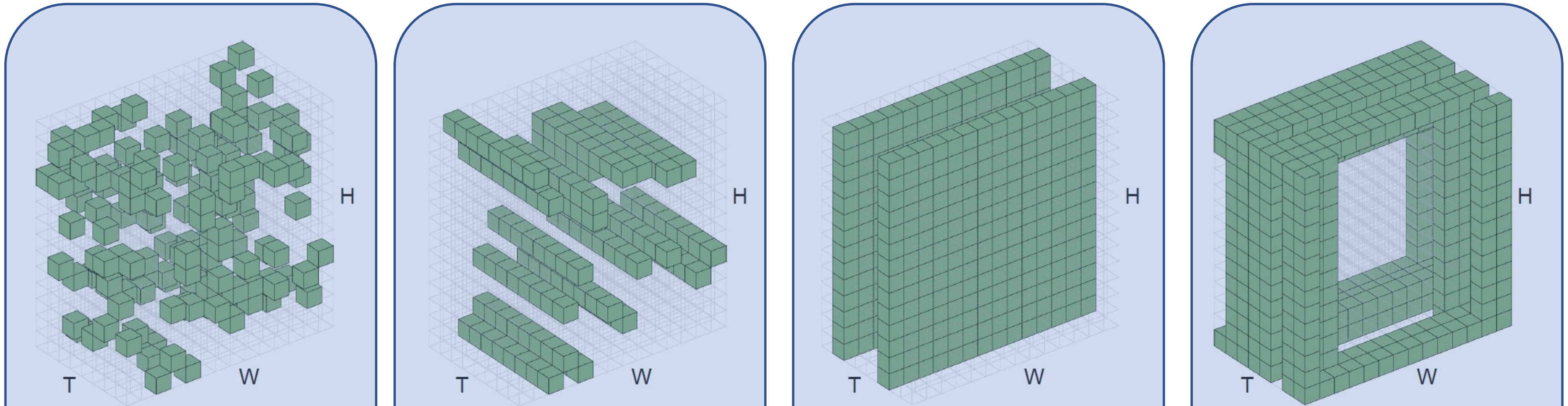
He *et al.*, Masked Autoencoders Are Scalable Vision Learners

Masked Autoencoders as spatiotemporal learners



- Masking of random patches in spacetime
- Encoder operates on the set of visible patches
- A small decoder on encoded patches and mask tokens reconstructs input
- Except for patch and positional embeddings, no inductive bias

Masking can be agnostic in spacetime



(a) agnostic, 90% (b) space-only, 90% (c) time-only, 75% (d) block, 75%

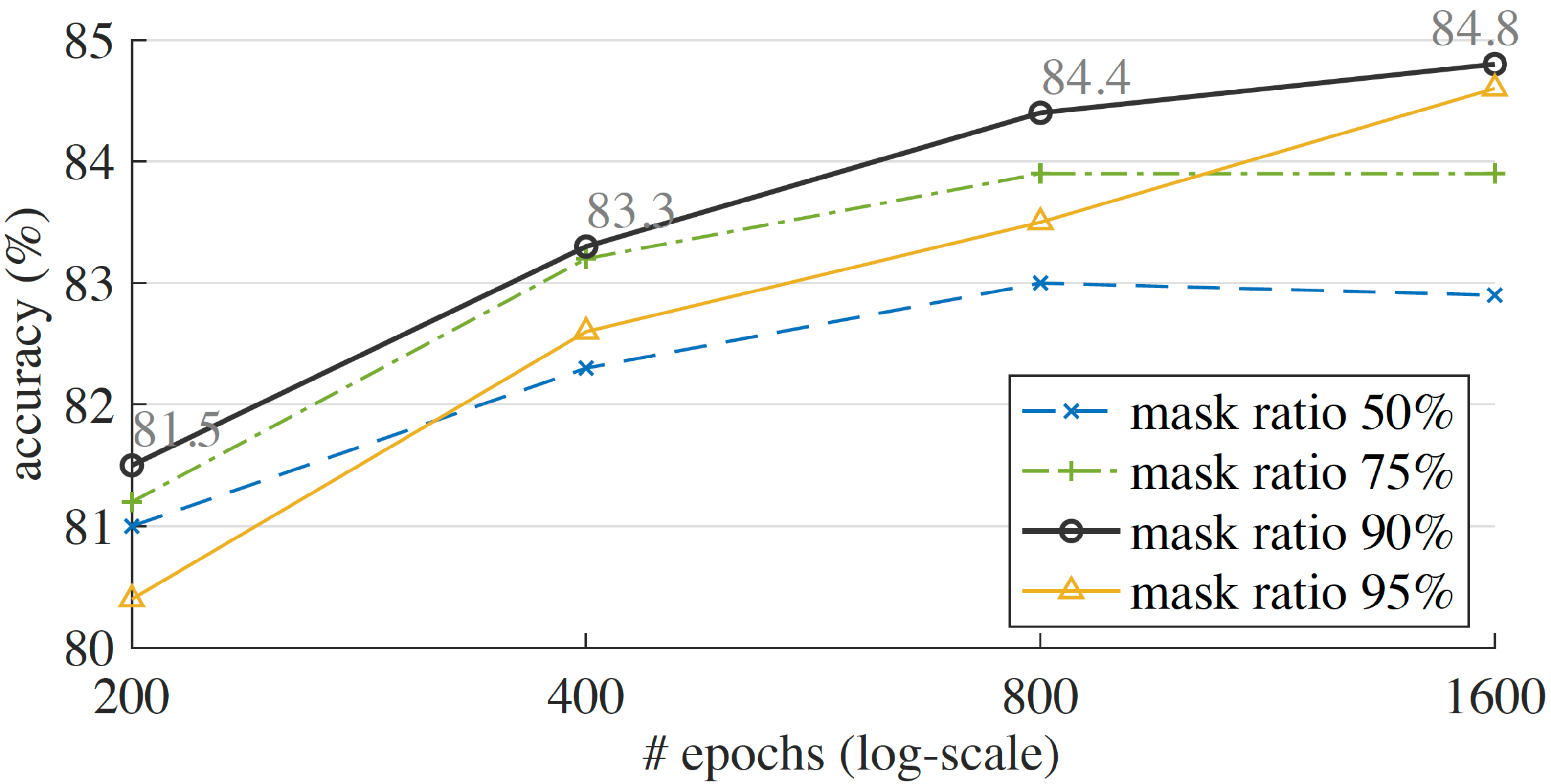
case	ratio	acc.
agnostic	90	**84.4**
space-only	90	83.5
time-only	75	79.1
block	75	83.2

(a) **Mask sampling**. See also Fig. 4. Random sampling that is spacetime-*agnostic* works the best.

- Task:
  Kinetics-400
  video classification

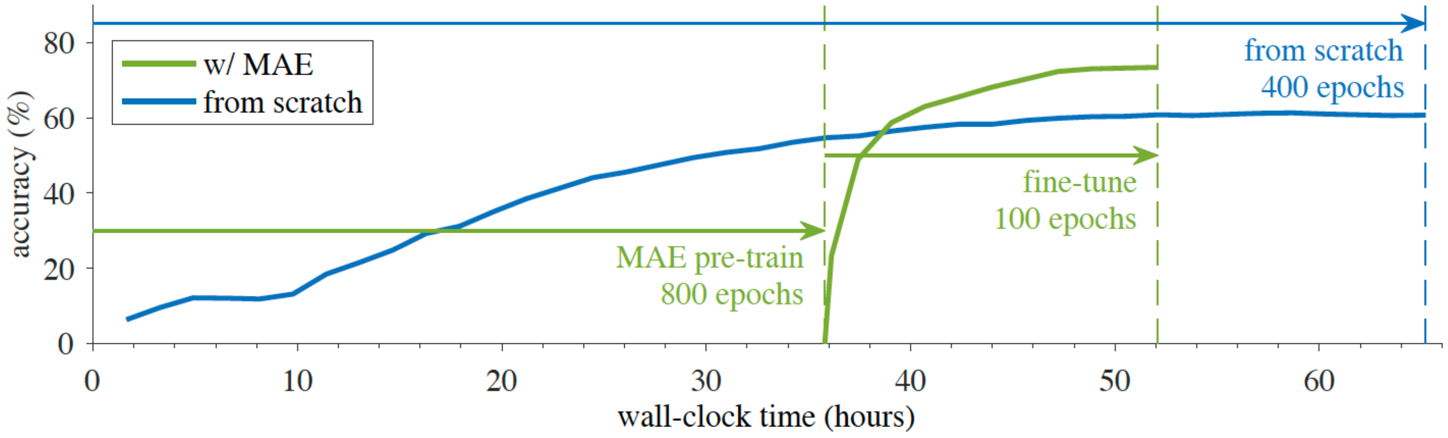
- Model: ViT-L
- Pre-train: 800 epochs
- Fine-tune: 100 epochs

Masking ratio can be extremely high



- For image classification, 75% is the optimal value, but for video 90% is considerably better

MAE is faster than pure supervised training



	scratch	MAE
1-view	60.7	**73.4** (+12.7)
multi-view	71.4	**84.4** (+13.0)

Figure 5: MAE pre-training plus fine-tuning is *much more accurate* and *faster* than training from scratch. Here the x-axis is the wall-clock training time (128 A100 GPUs), and the y-axis is the 1-view accuracy on Kinetics-400 validation. The table shows the final accuracy. The model is ViT-L.

Influence of data scale and curation

pre-train set	# pre-train data	pre-train method	K400	AVA	SSv2
-	-	none (from scratch)	71.4	-	-
K400	240k	supervised	-	21.6	55.7

Table 3: **Influence of pre-training data**, evaluated on K400, AVA, and SSv2 as the downstream tasks.

MAE visualizations

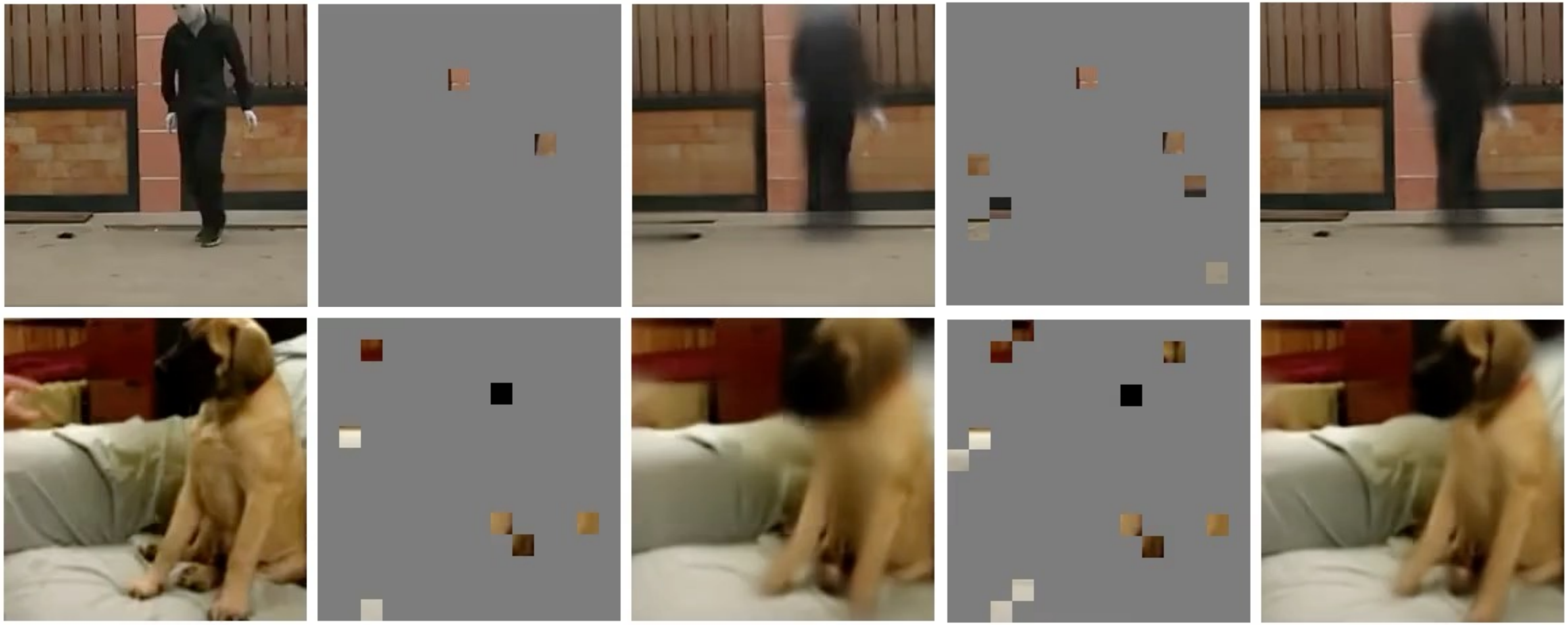
original | input 98% masked | output 98% | input 95% masked | output 95%

MAE visualizations

original | input 98% masked | output 98% | input 95% masked | output 95%

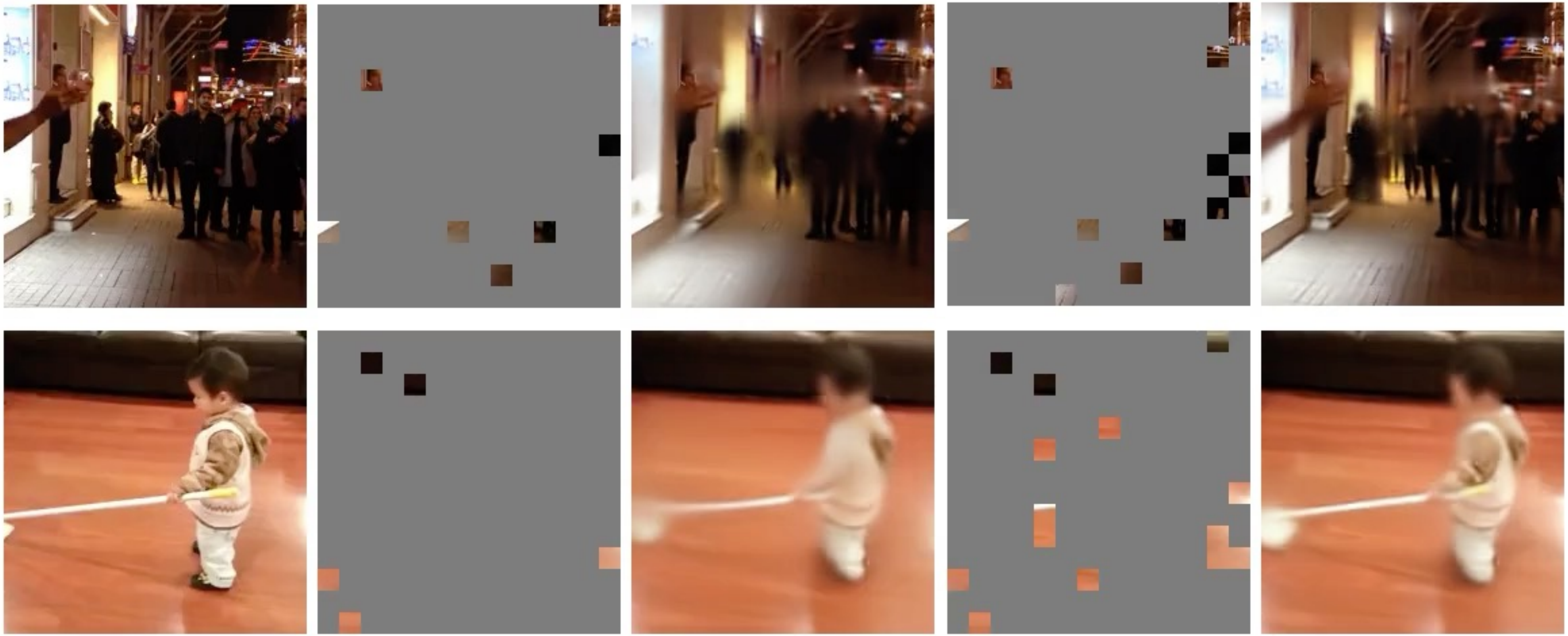
MAE visualizations

original | input 98% masked | output 98% | input 95% masked | output 95%



MAE visualizations

original | input 98% masked | output 98% | input 95% masked | output 95%



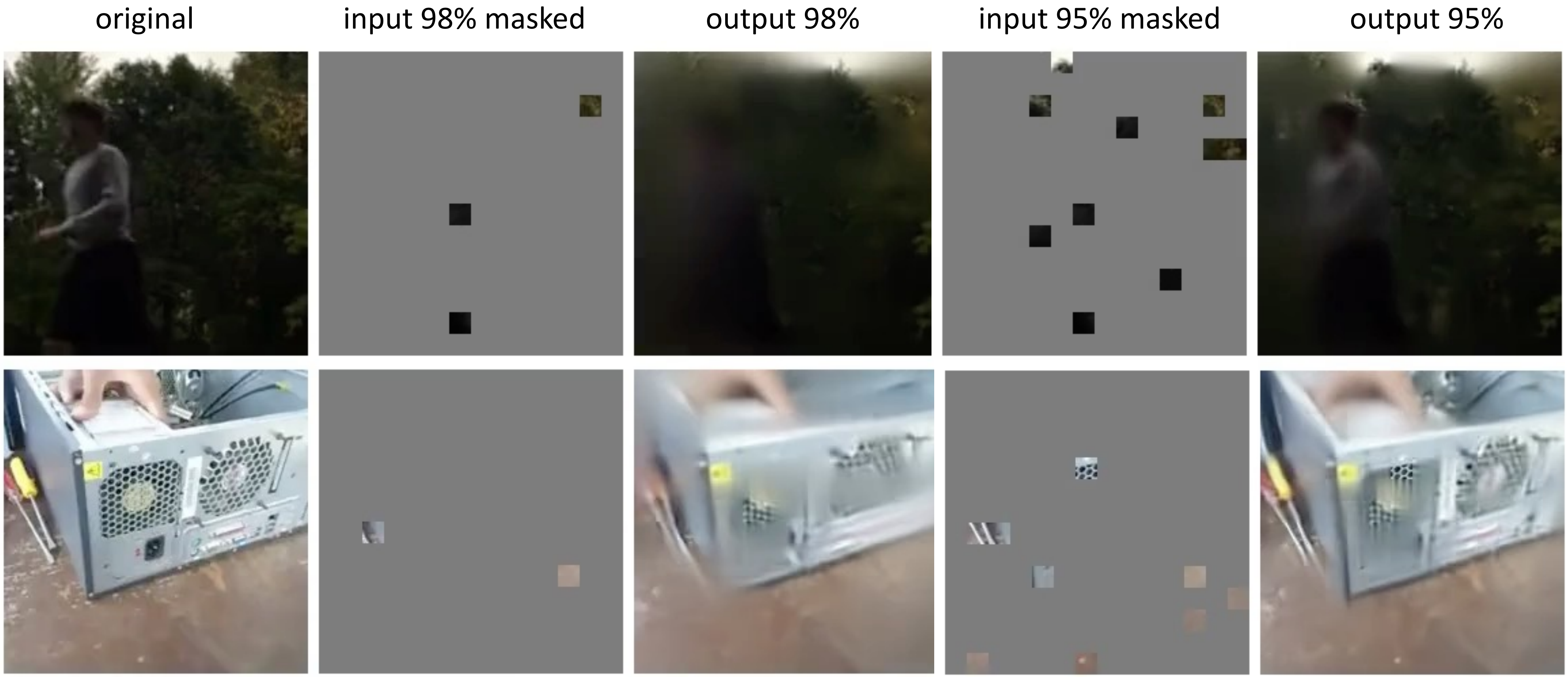
MAE visualizations

original | input 98% masked | output 98% | input 95% masked | output 95%

MAE visualizations

original | input 98% masked | output 98% | input 95% masked | output 95%

MAE visualizations



Masked Feature Prediction for Self-Supervised Visual Pre-Training

Chen Wei[*,1,2], Haoqi Fan[1], Saining Xie[1], Chao-Yuan Wu[1], Alan Yuille[2], Christoph Feichtenhofer[*,1]
[1]Meta AI, FAIR, [2]Johns Hopkins University

In CVPR 2022

github.com/facebookresearch/SlowFast

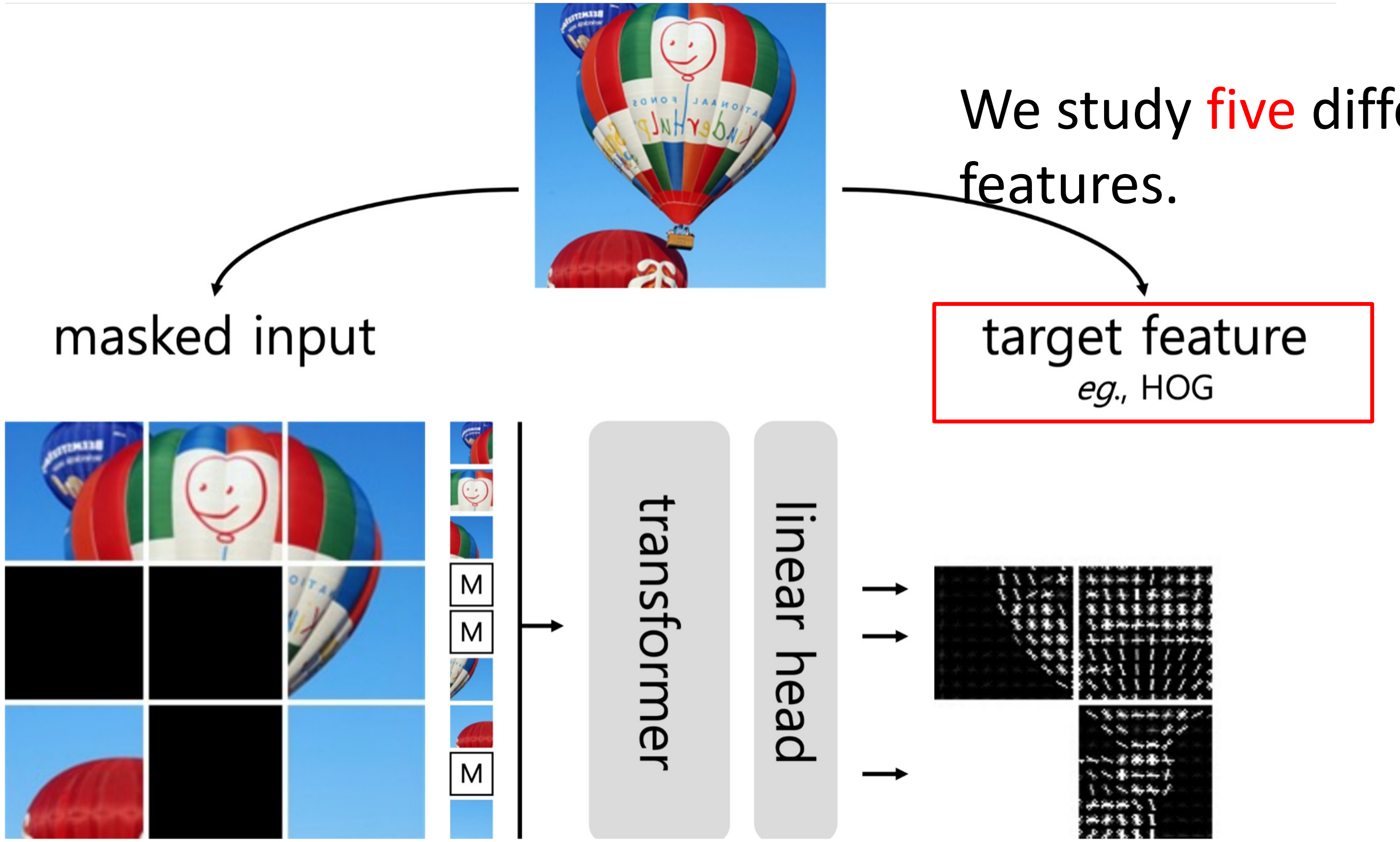
Language *vs.* Vision

- Language
  - sparse, discrete, semantic-rich
  - natural word tokens

- Vision
  - dense, continuous, high-dimensional
  - mimicking language: visual words/codebook?

Masked Feature Prediction

We study five different types of features.



regress the masked patches

Feature #1: pixel colors

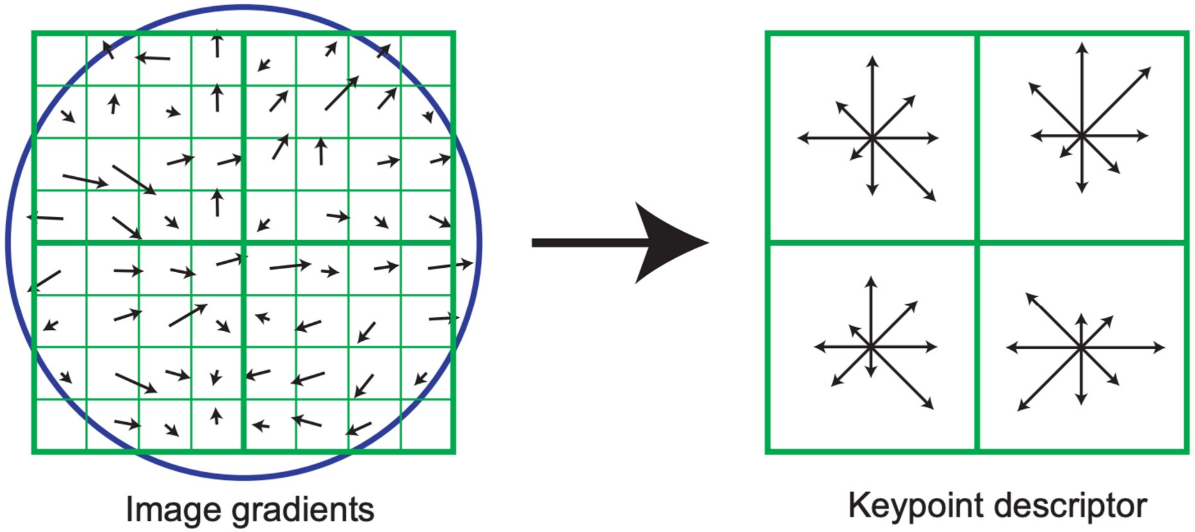
- RGB raw pixels
  - A small gain
  - trivial local statistics and high-frequency details

feature type	one-stage	variant	arch.	param.	epoch†	top-1
scratch	-	DeiT [84]	-	-	-	81.8
pixel colors	✓	RGB	-	-	-	82.5

+0.7

Feature #2: HOG

- Histogram of Oriented Gradients
  - popular in 2000s
  - invariance to geometry and photometric change (to some extent)
  - fast to compute with pytorch and GPU



from SIFT paper

Feature #2: HOG

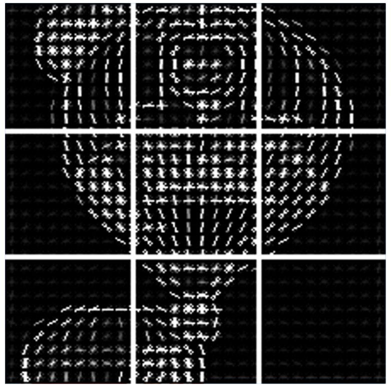
Input image

Histogram of Oriented Gradients

from scikit-image

Feature #2: HOG

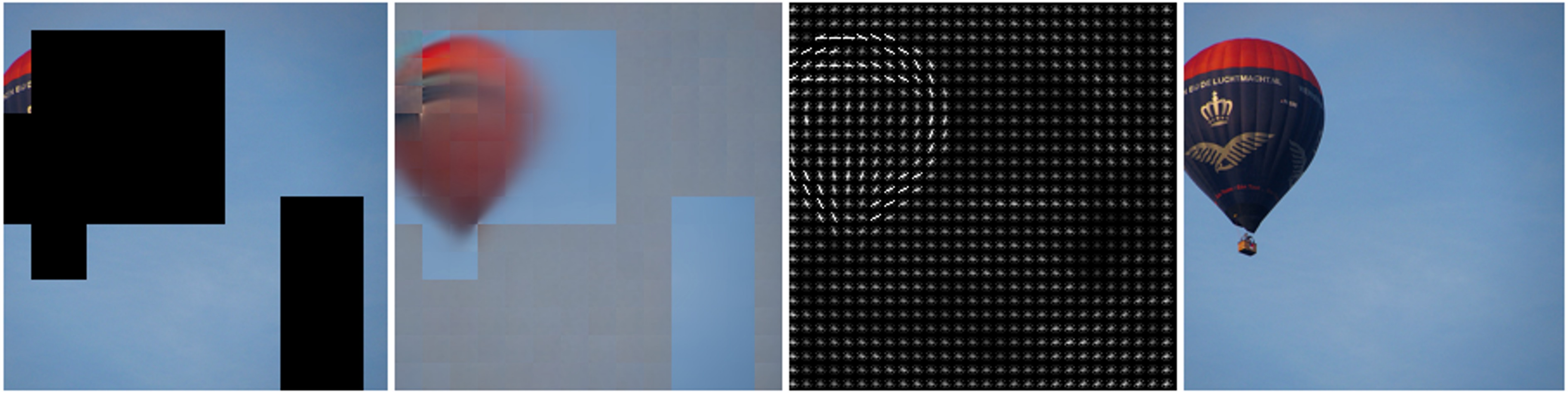
- Histogram of Oriented Gradients
  - invariance helps!



feature type	one-stage	variant	arch.	param.	epoch†	top-1
scratch	-	DeiT [84]	-	-	-	81.8
pixel colors	✓	RGB	-	-	-	82.5
image descriptor	✓	HOG [22]	-	-	-	**83.6**

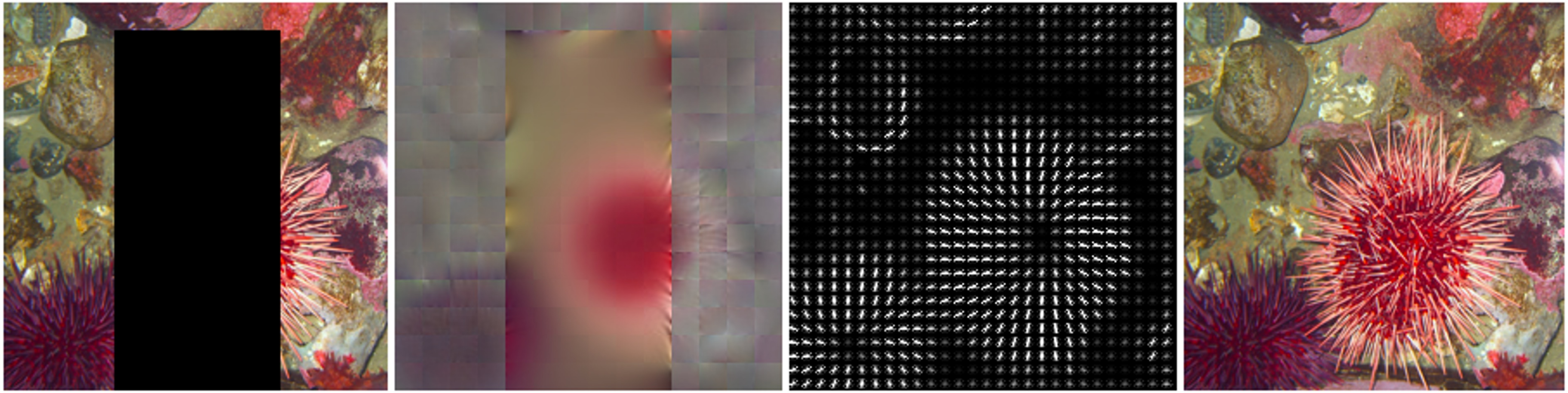
+1.8

Pixel *vs.* HOG: Color Ambiguity



pixel: large loss penalty because of unmatched color

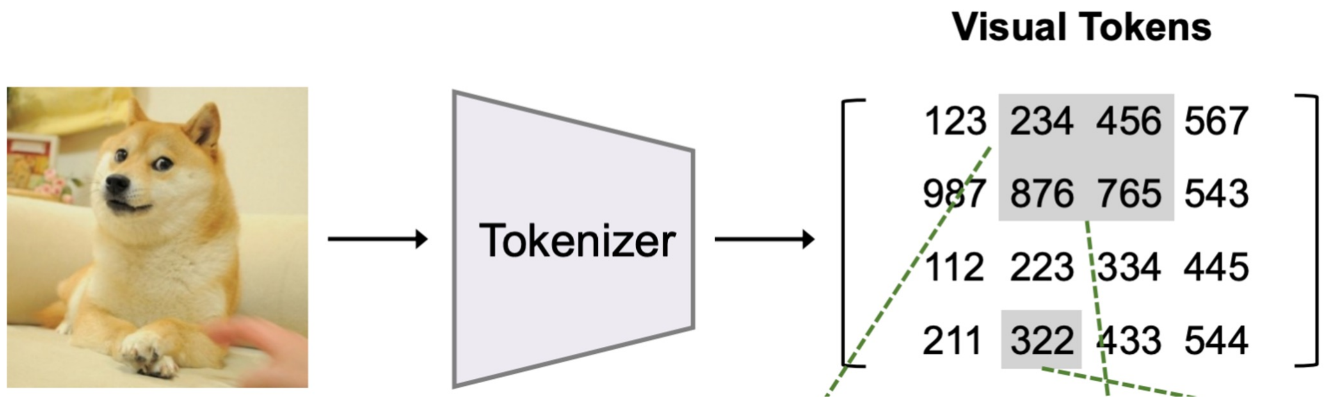
Pixel *vs.* HOG: Texture Ambiguity



HOG: captures major edge directions

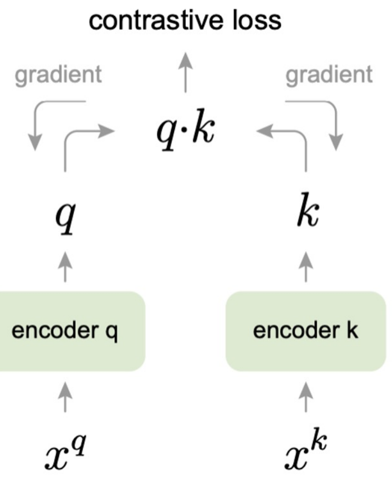
Feature #3: token

- discrete VAE token
  - patch clustering
  - BEiT



feature type	one-stage	variant	arch.	param.	epoch†	top-1
scratch	-	DeiT [84]	-	-	-	81.8
pixel colors	✓	RGB	-	-	-	82.5
image descriptor	✓	HOG [22]	-	-	-	**83.6**
dVAE token	✗	DALL-E [73]	dVAE	54	1199	82.8

Feature #4: deep features



- unsupervised deep features
  - contrastive unsupervised methods
  - work better than others

feature type	one-stage	variant	arch.	param.	epoch†	top-1
scratch	-	DeiT [84]	-	-	-	81.8
pixel colors	✓	RGB	-	-	-	82.5
image descriptor	✓	HOG [22]	-	-	-	**83.6**
dVAE token	✗	DALL-E [73]	dVAE	54	1199	82.8
unsupervised feature	✗	MoCo v2 [16]	ResNet50	23	800	83.6
unsupervised feature	✗	MoCo v3 [18]	ViT-B	85	600	83.9
unsupervised feature	✗	DINO [9]	ViT-B	85	1535	**84.0**

+2.2

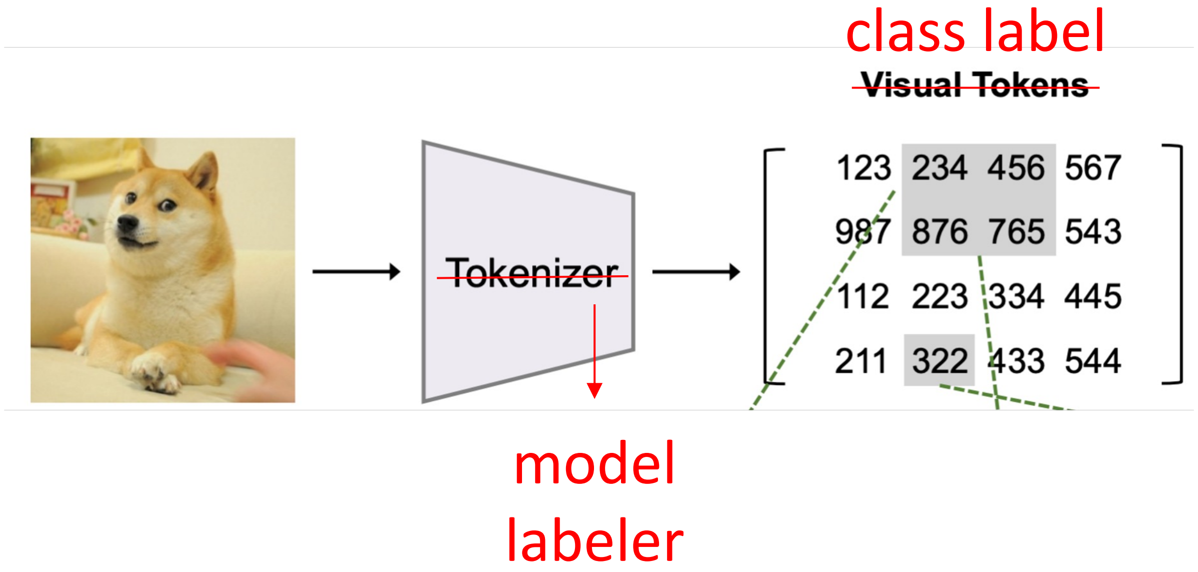
Feature #4: deep features

- supervised deep features
  - more labels, lower top-1
  - ResNet50 helps, ViT-B does not

feature type	one-stage	variant	arch.	param.	epoch†	top-1
scratch	-	DeiT [84]	-	-	-	81.8
pixel colors	✓	RGB	-	-	-	82.5
image descriptor	✓	HOG [22]	-	-	-	**83.6**
dVAE token	✗	DALL-E [73]	dVAE	54	1199	82.8
unsupervised feature	✗	MoCo v2 [16]	ResNet50	23	800	83.6
unsupervised feature	✗	MoCo v3 [18]	ViT-B	85	600	83.9
unsupervised feature	✗	DINO [9]	ViT-B	85	1535	**84.0**
supervised feature	✗	pytorch [67]	ResNet50	23	90	82.6
supervised feature	✗	DeiT [84]	ViT-B	85	300	81.9

Feature #5: pseudo label

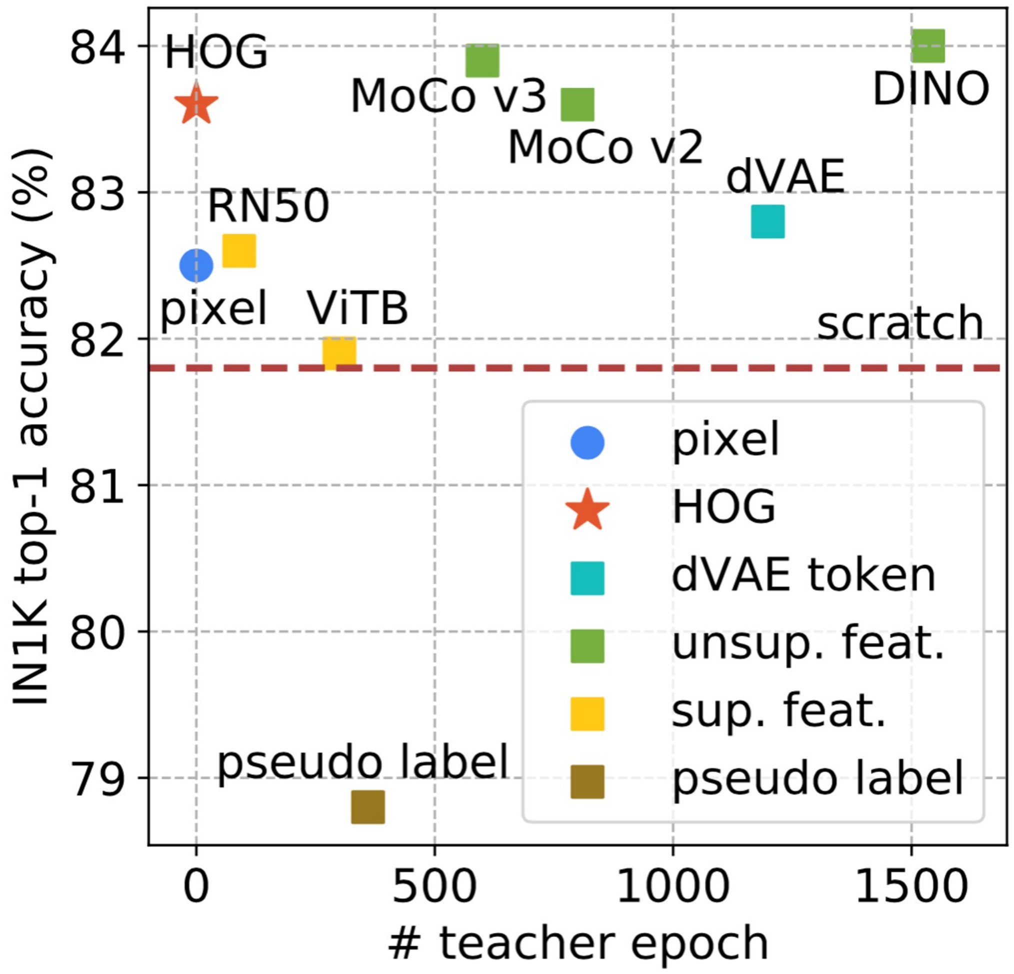
- pseudo class label for each patch
  - labeled by a 86.5% supervised model
  - but results in a huge drop



feature type	one-stage	variant	arch.	param.	epoch†	top-1
scratch	-	DeiT [84]	-	-	-	81.8
pixel colors	✓	RGB	-	-	-	82.5
image descriptor	✓	HOG [22]	-	-	-	**83.6**
dVAE token	✗	DALL-E [73]	dVAE	54	1199	82.8
unsupervised feature	✗	MoCo v2 [16]	ResNet50	23	800	83.6
unsupervised feature	✗	MoCo v3 [18]	ViT-B	85	600	83.9
unsupervised feature	✗	DINO [9]	ViT-B	85	1535	**84.0**
supervised feature	✗	pytorch [67]	ResNet50	23	90	82.6
supervised feature	✗	DeiT [84]	ViT-B	85	300	81.9
pseudo-label	✗	Token Labeling [50]	NFNet-F6	438	360	78.8

-3.0

Masked Feature Prediction



ImageNet-1K Fine-Tuning

norm.	none	ℓ_1	ℓ_2
top-1	82.2	82.8	**83.6**

(a) **Contrast normalization.**

channel	gray	rgb	opp.
top-1	83.2	**83.6**	83.5

(b) **Color channel.**

#bins	6	9	12
top-1	83.4	**83.6**	83.5

(c) **Orientation bins.**

cell size	4×4	8×8	16×16
top-1	83.2	**83.6**	83.2

(d) **Spatial cell size.**

pre-train	extra data	extra model	ViT-B	ViT-L
scratch [84]	-	-	81.8	81.5
supervised$_{384}$ [27]	IN-21K	-	84.0	85.2
MoCo v3 [18]	-	momentum ViT	83.2	84.1
DINO [9]	-	momentum ViT	82.8	-
BEiT [2]	DALL-E	dVAE	83.2	85.2
MaskFeat (w/ HOG)	-	-	**84.0**	**85.7**

+4.2

ImageNet val accuracy

Masked Autoencoders that Listen

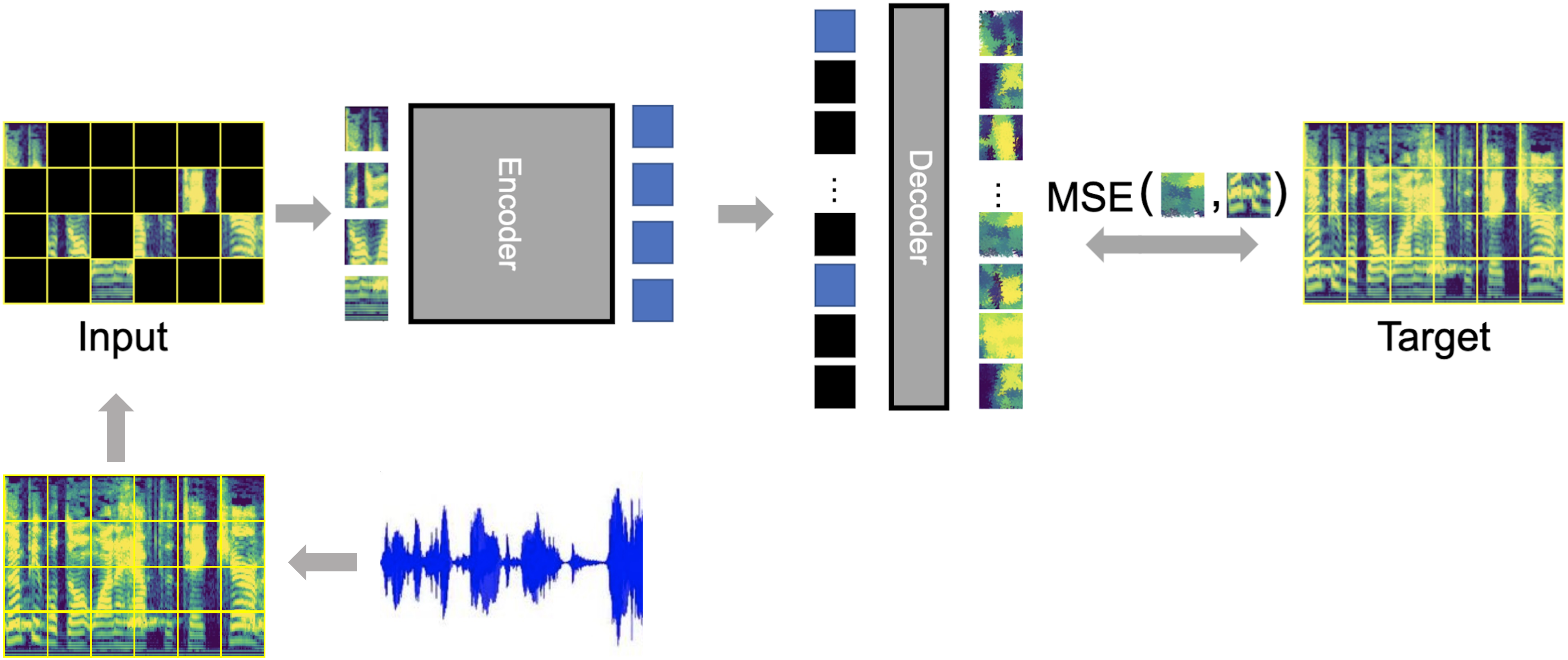
Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski
Michael Auli, Wojciech Galuba, Florian Metze, Christoph Feichtenhofer

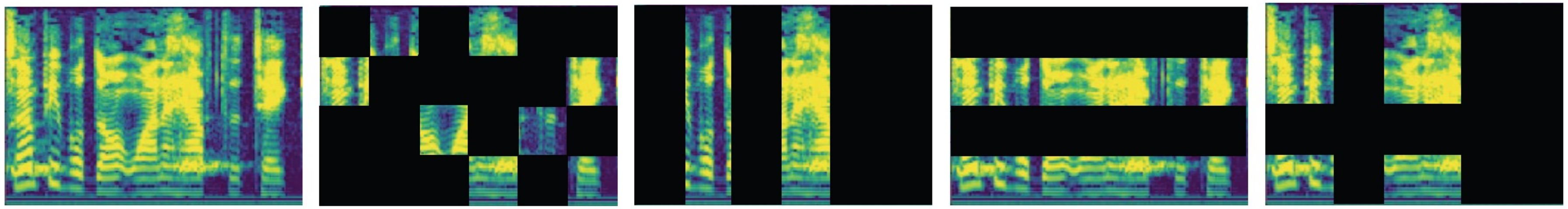
Meta AI, FAIR

In NeurIPS 2022

github.com/facebookresearch/AudioMAE

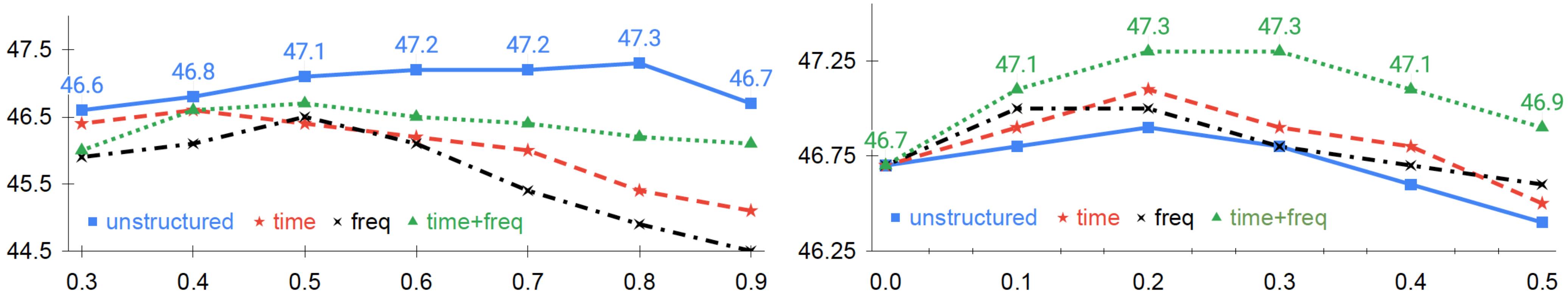
Audio-MAE





(a) Original (b) Unstructured (c) Time (d) Frequency (e) Time+frequency

Figure 2: Masking strategies for Audio-MAE.



(a) Pre-training masking (b) Fine-tuning masking

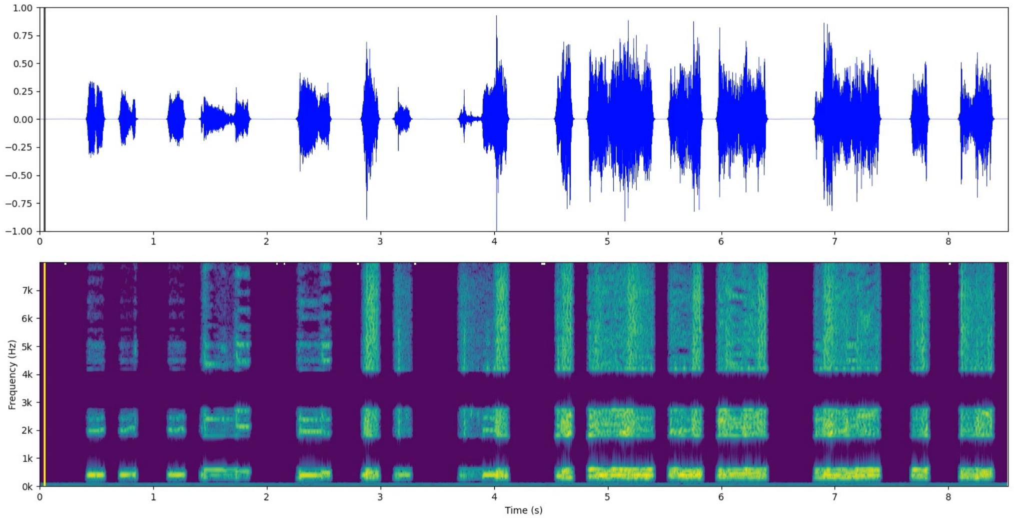
Figure 4: **Masking strategy**. A *higher* ratio and *unstructured* masking (random) is preferred in audio pre-training. For fine-tuning, a *lower* ratio and *structured* masking (time+frequency) is better. The y-axes are mAP on AS-2M and the x-axes are masking ratio.

Comparison to state-of-the-art

Model	Backbone	PT-Data	AS-20K	AS-2M	ESC-50	SPC-2	SPC-1	SID
No pre-training								
ERANN [57]	CNN	-	-	45.0	89.2	-	-	-
PANN [58]	CNN	-	27.8	43.1	83.3	61.8	-	-
In-domain self-supervised pre-training								
wav2vec 2.0 [33]	Transformer	LS	-	-	-	-	96.2*	75.2*
HuBERT [35]	Transformer	LS	-	-	-	-	96.3*	81.4*
Conformer [37]	Conformer	AS	-	41.1	88.0	-	-	-
SS-AST [18]	ViT-B	AS+LS	31.0	-	88.8	98.0	96.0	64.3
Concurrent MAE-based works								
MaskSpec [43]	ViT-B	AS	32.3	47.1	89.6	97.7	-	-
MAE-AST [38]	ViT-B	AS+LS	30.6	-	90.0	97.9	95.8	63.3
Audio-MAE (global)	ViT-B	AS	36.6±.11	46.8±.06	93.6±.11	**98.3**±.06	**97.6**±.06	94.1±.06
Audio-MAE (local)	ViT-B	AS	**37.1**±.06	**47.3**±.06	**94.1**±.10	**98.3**±.06	96.9±.00	**94.8**±.11
Out-of-domain supervised pre-training								
PSLA [30]	EffNet [59]	IN	31.9	44.4	-	96.3	-	-
AST [10]	DeiT-B	IN	34.7	45.9	88.7	98.1	95.5	41.1
MBT [11]	ViT-B	IN-21K	31.3	44.3	-	-	-	-
HTS-AT [29]	Swin-B	IN	-	47.1	97.0†	98.0	-	-
PaSST [28]	DeiT-B	IN	-	47.1	96.8†	-	-	-

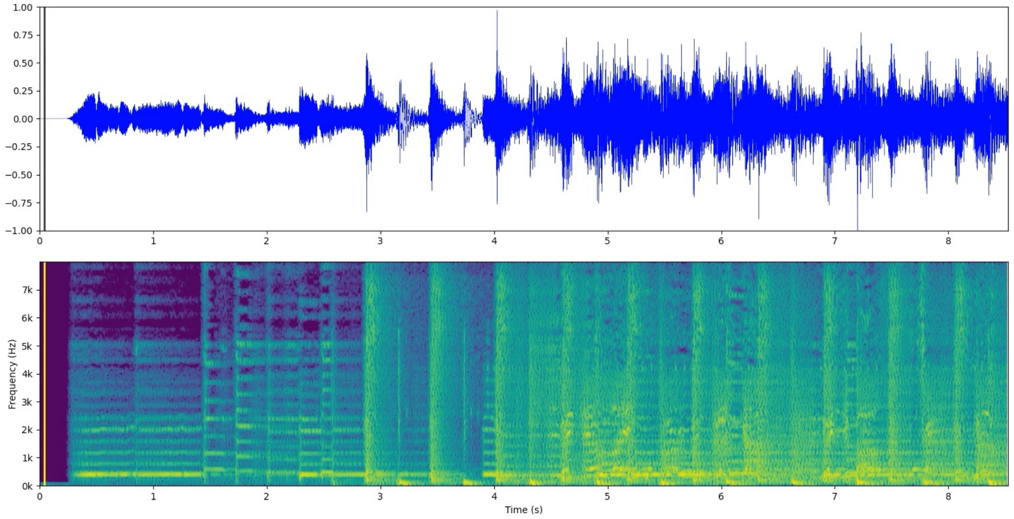
Audio-MAE music sample, *structured masking*

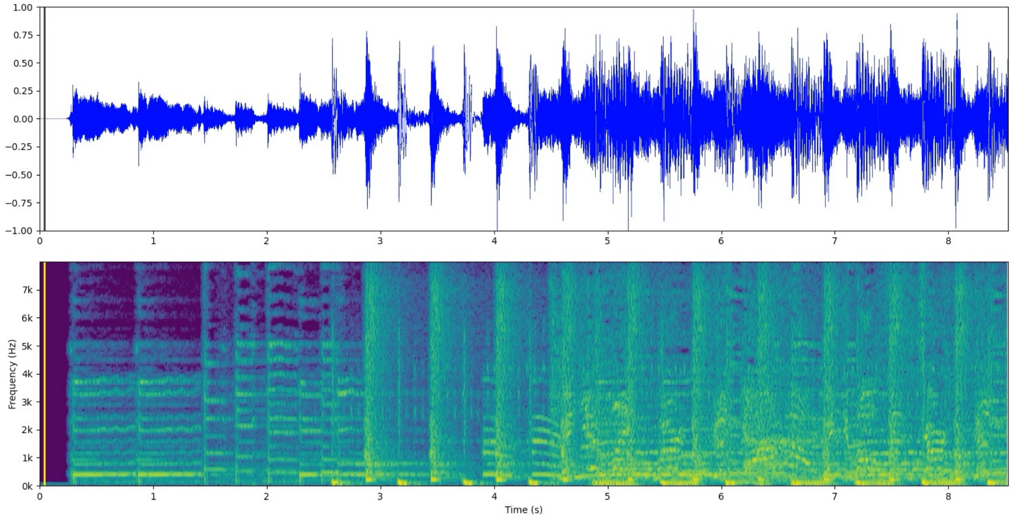
masked 80%



original

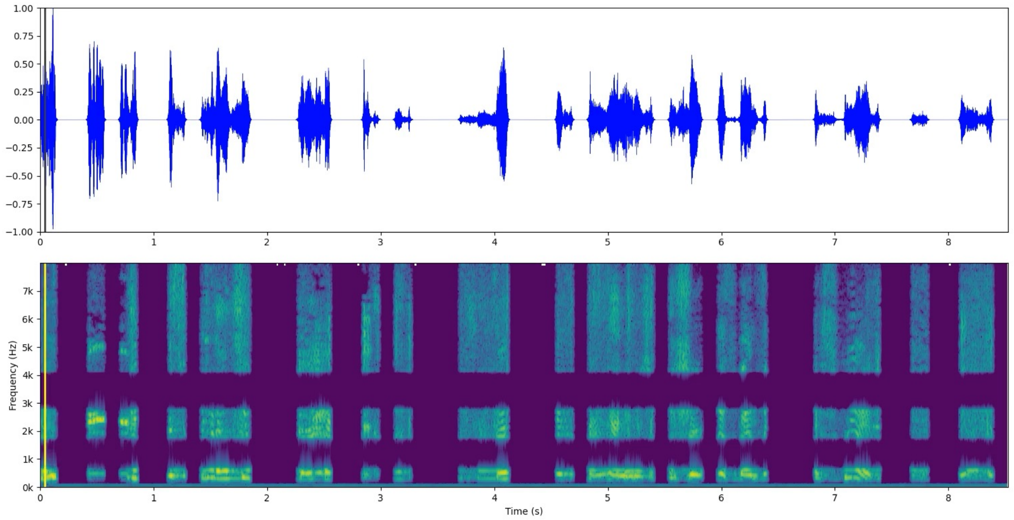
reconstruction output





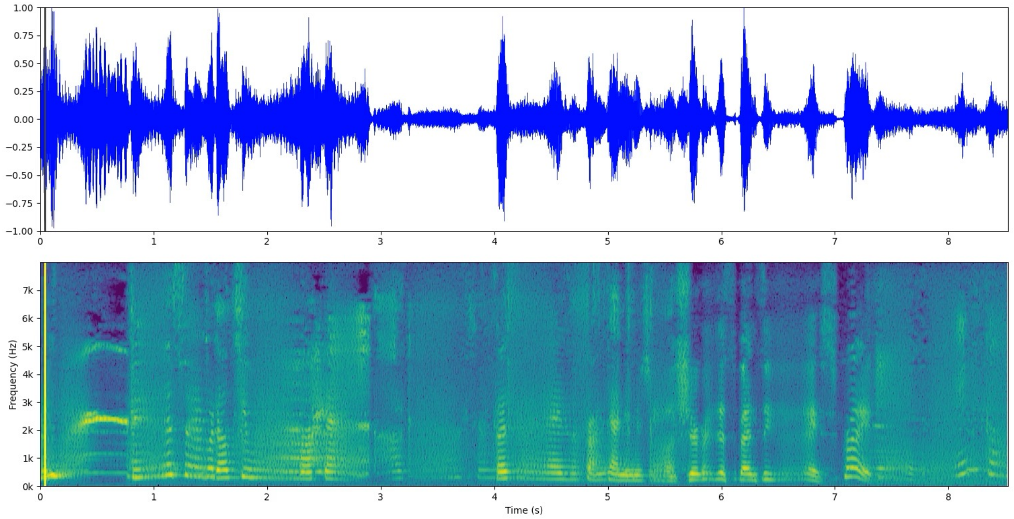
Audio-MAE speech sample, *structured masking*

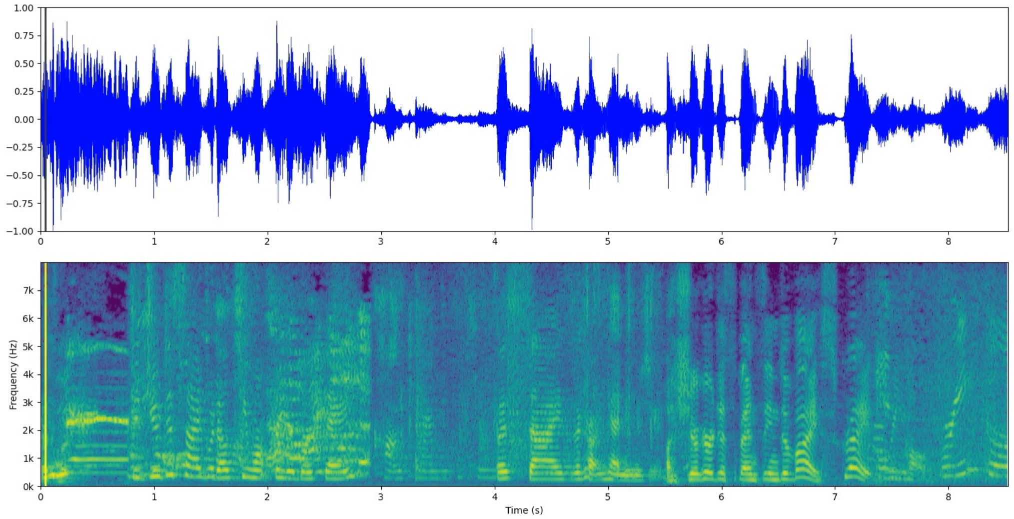
masked 80%



original

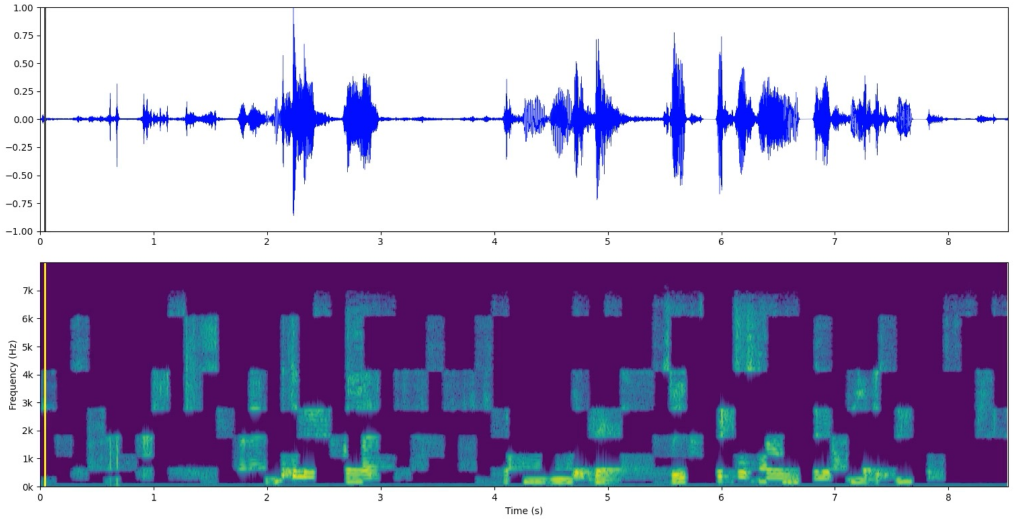
reconstruction output





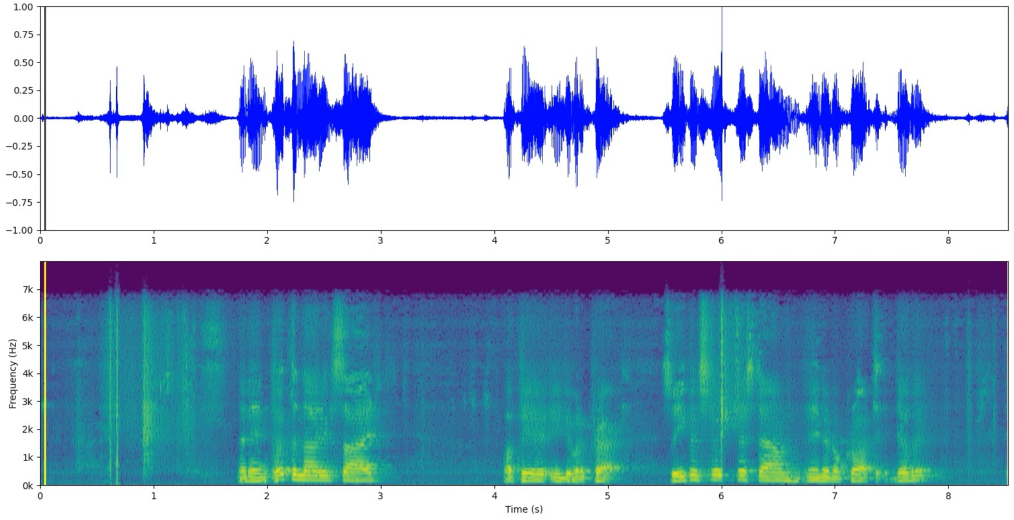
Audio-MAE speech sample

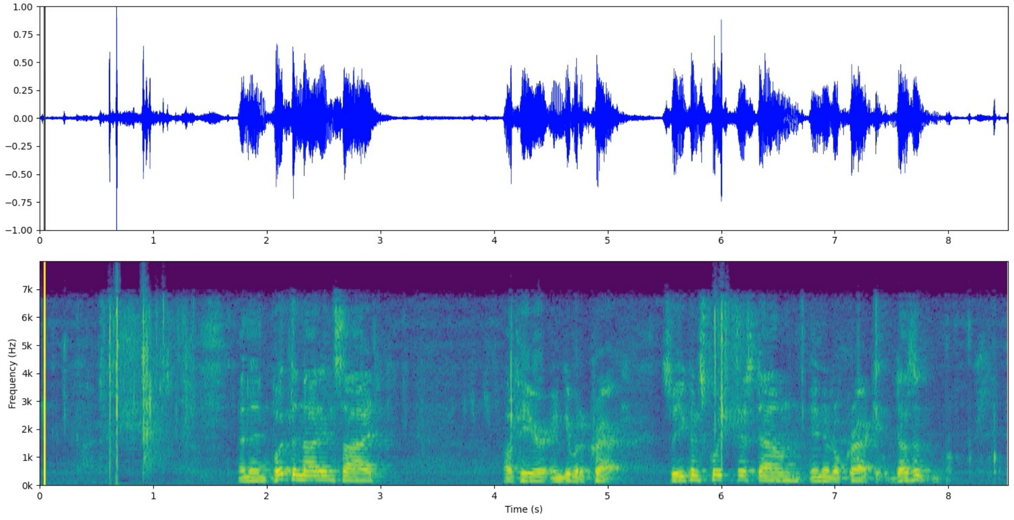
masked 80%



original

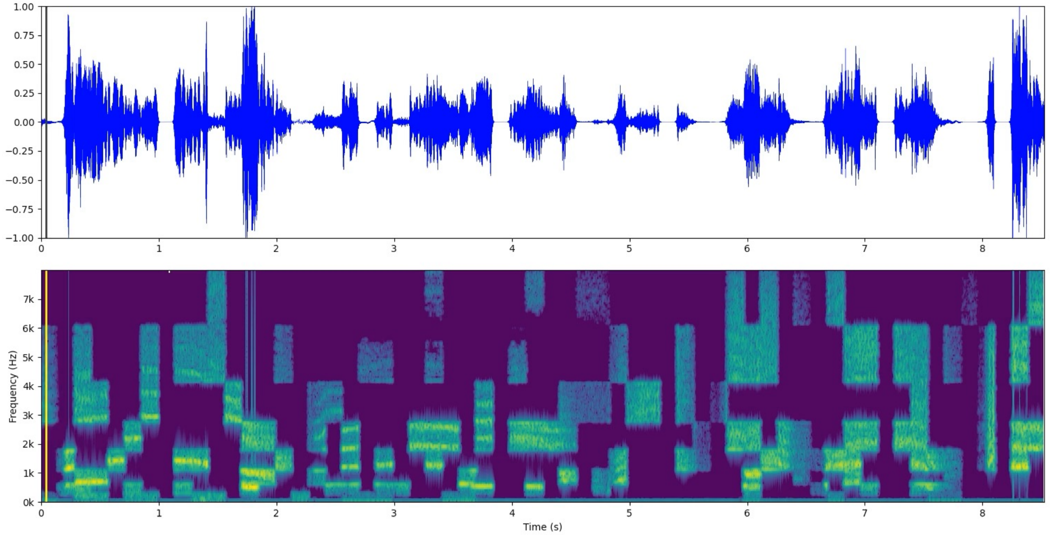
reconstruction output





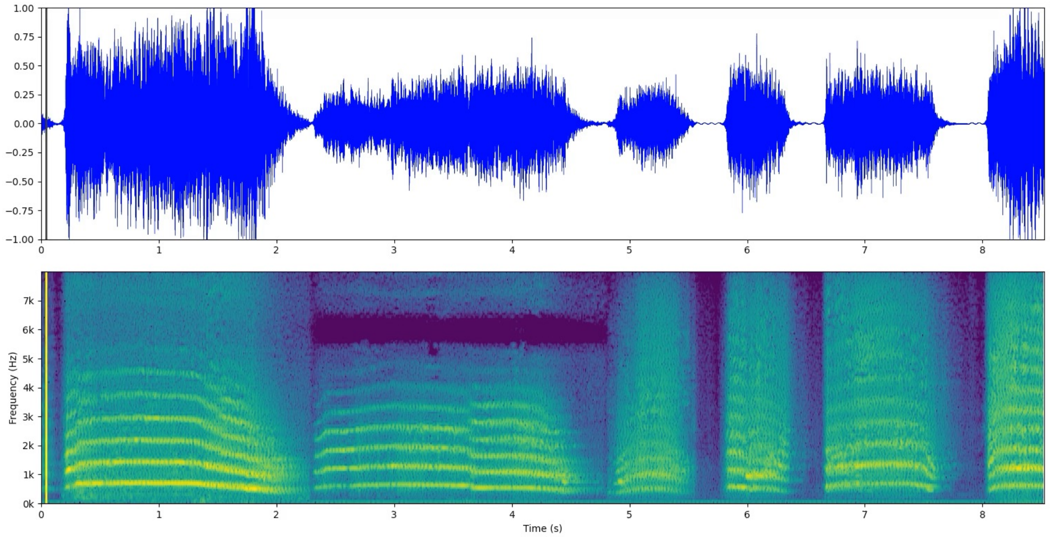
Audio-MAE misc sound sample

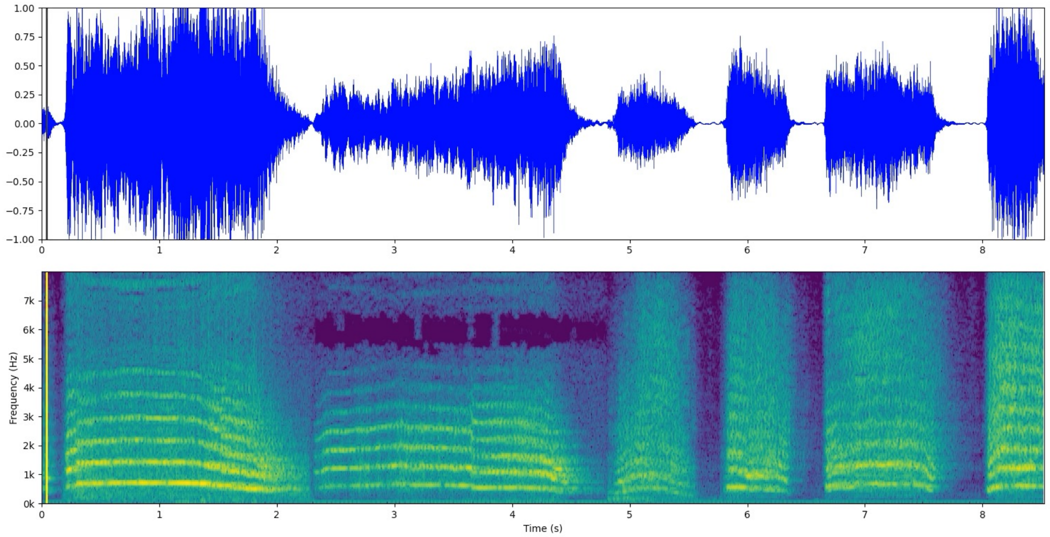
masked 70%



original

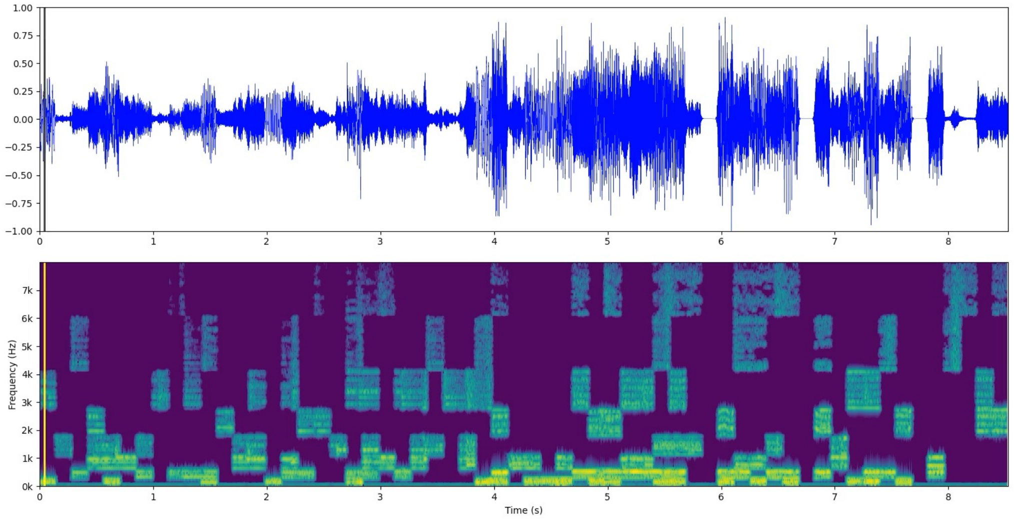
reconstructio
output





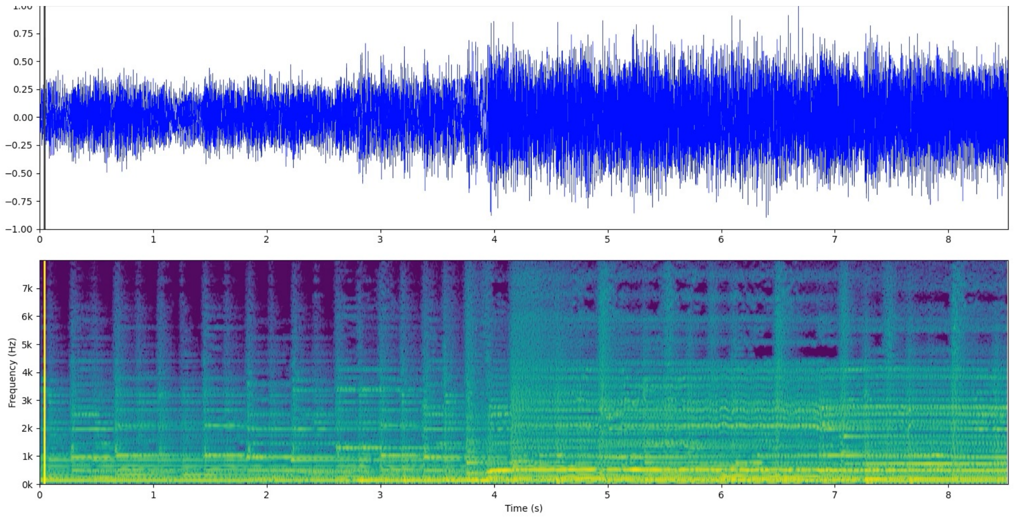
Audio-MAE music sample

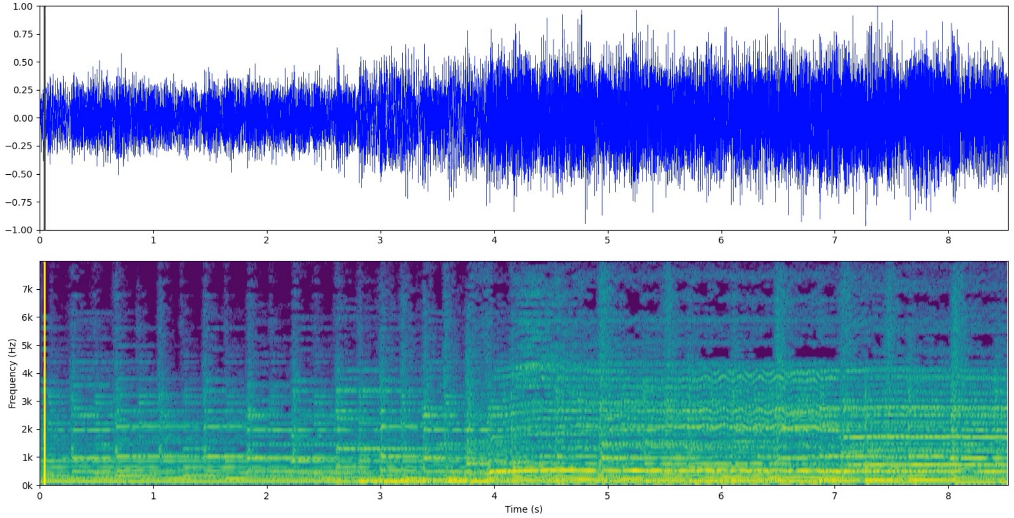
masked 80%



original

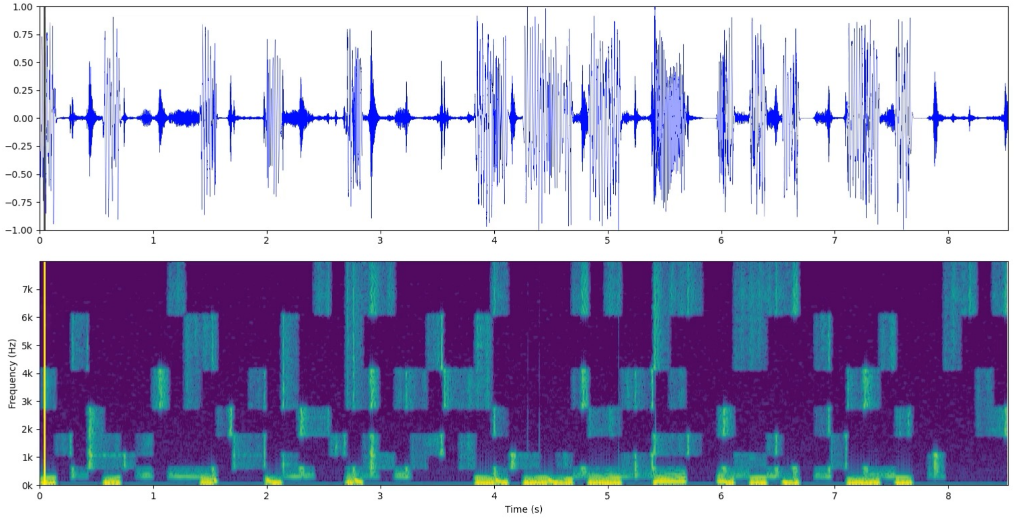
reconstruction output





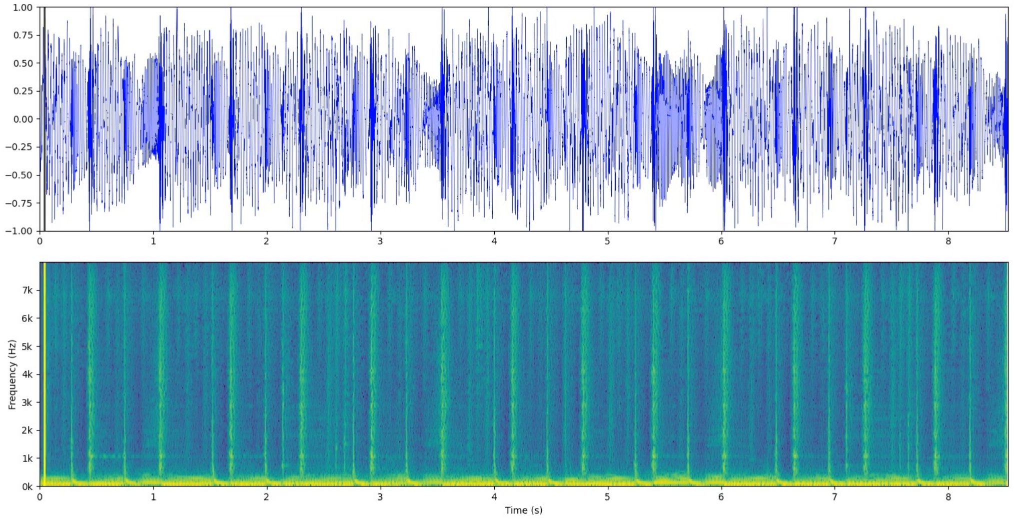
Audio-MAE music sample

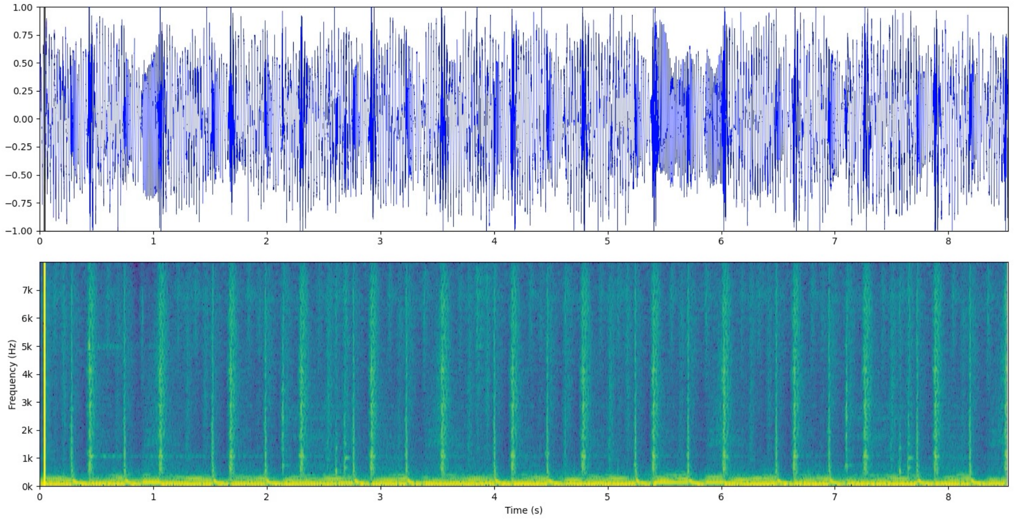
masked 80%



original

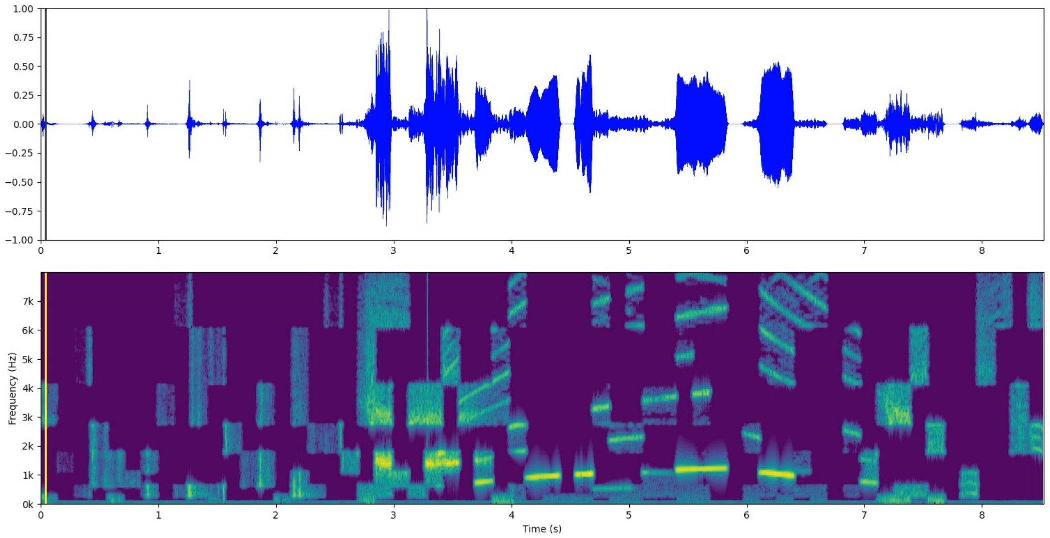
reconstructio
output





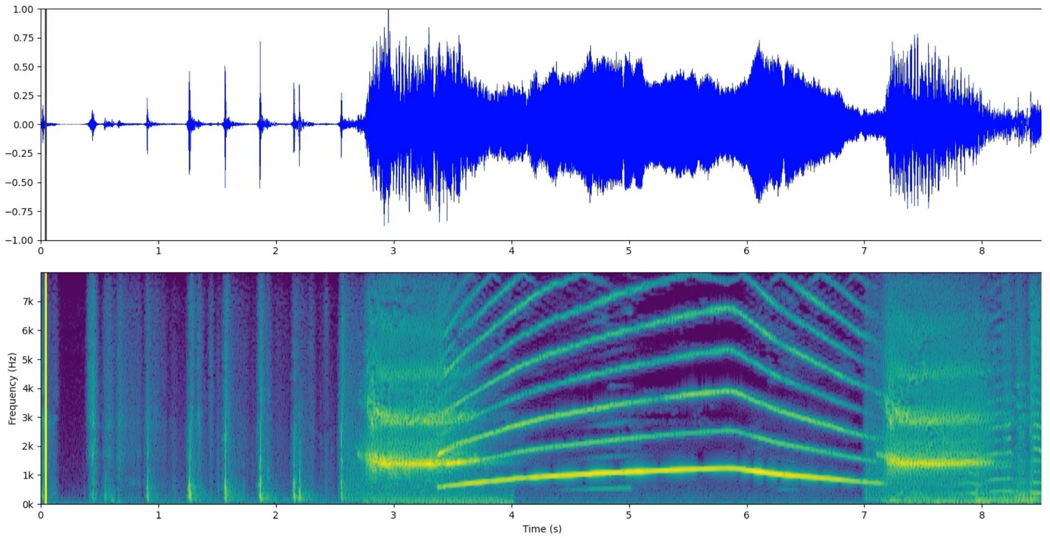
Audio-MAE event sound sample

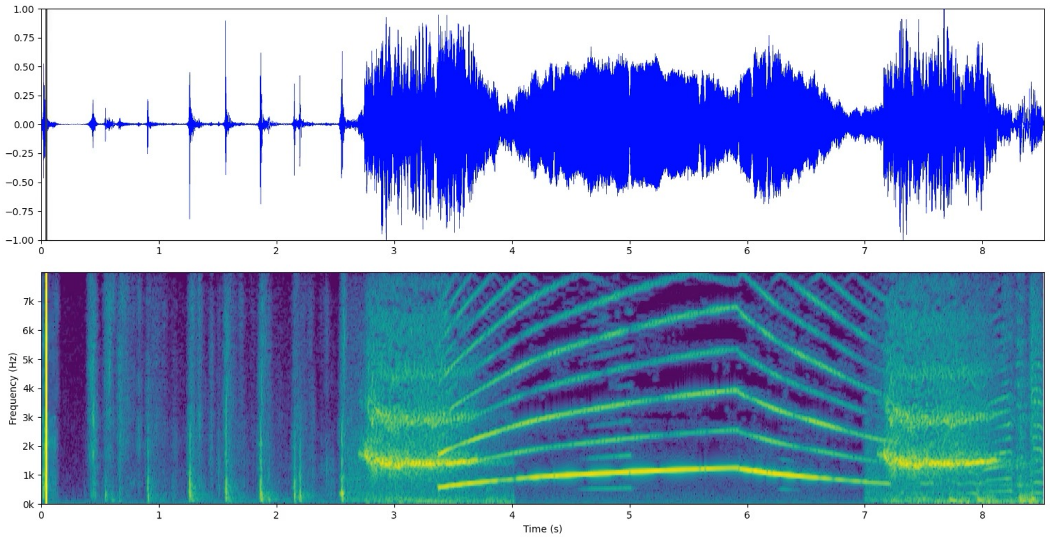
masked 80%



original

reconstruction output





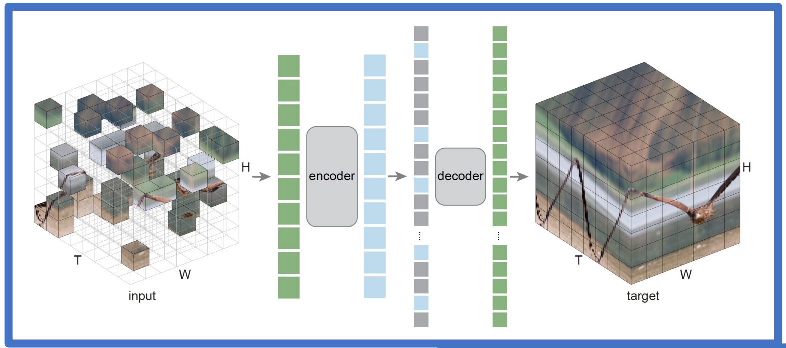
MAE Works Particularly on Video Because…

- Videos datasets are (relatively) small in terms of #videos
  - Low diversity - easy to overfit if training from scratch
  - Image pre-training for video has a domain gap
  - Directly pre-training on video is advantageous
- Videos are visually richer than images
  - Natural and abundant views of one object through time
  - One class label can not fully capture it – low "**label density**"
  - MAE directly learns to reconstruct both appearance and motion
- Masked autoencoding is general and the optimal masking strategy depends on the nature of the data (text, audio, image, video, … etc.)

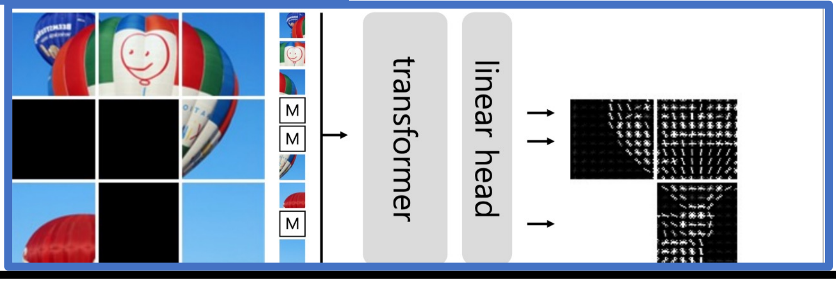
Summary: Unsupervised learning from video

- Video allows learning from spatiotemporal associations (across modalities)
- Offers to learn temporal prediction of appearance/shape, motion, as well as causality

1. Video MAE



2. MaskFeat



3. Audio MAE