MAViL: Masked Audio-Video Learners

Po-Yao Huang, Vasu Sharma, Hu Xu, Chaitanya Ryali, Haoqi Fan, Yanghao Li, Shang-Wen Li, Gargi Ghosh, Jitendra Malik, Christoph Feichtenhofer

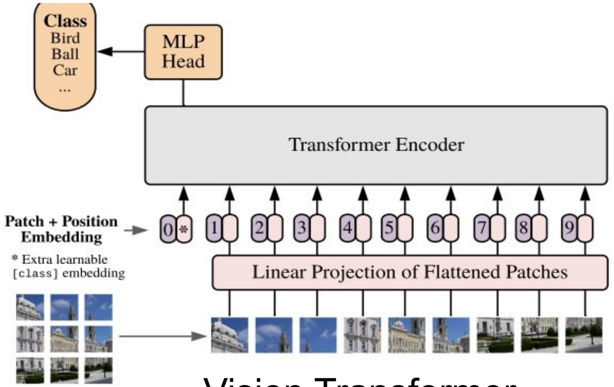
FAIR Labs 2023

Outline

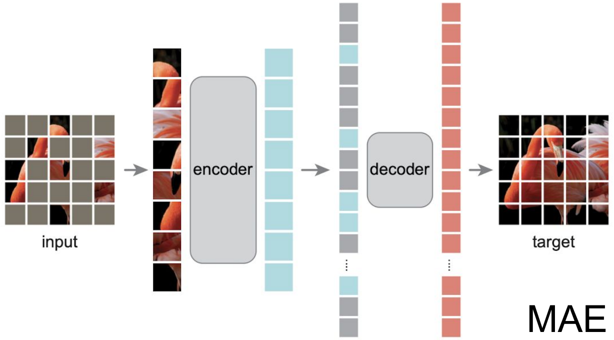
- Background: A unifying trend in single-modal self-supervised learning (SSL)
- MAViL: Masked Audio-Video Learners
    - Prior work in audio-video SSL
    - Motivation: predicting ~~hetereogenous~~ homogenous (aligned) and ~~raw~~ contextualized targets.
- Experimental Results
- Conclusion

A unifying trend across CV, NLP, Audio

ImageNet

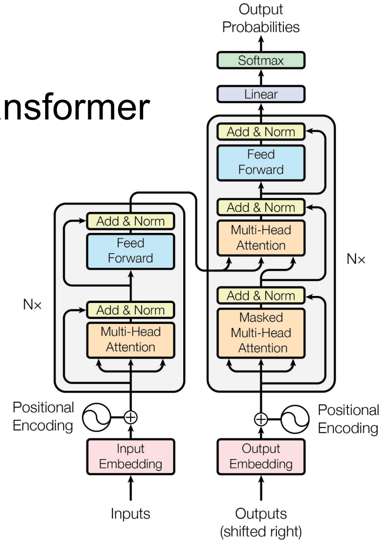


Vision Transformer

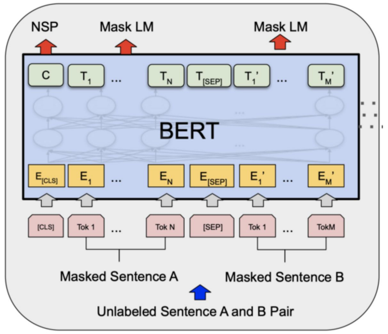


MAE

Transformer

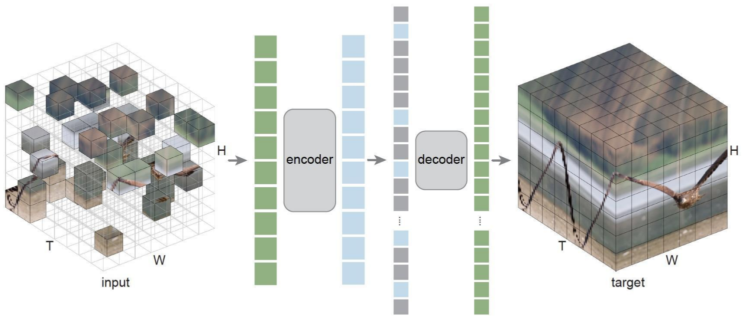


BERT

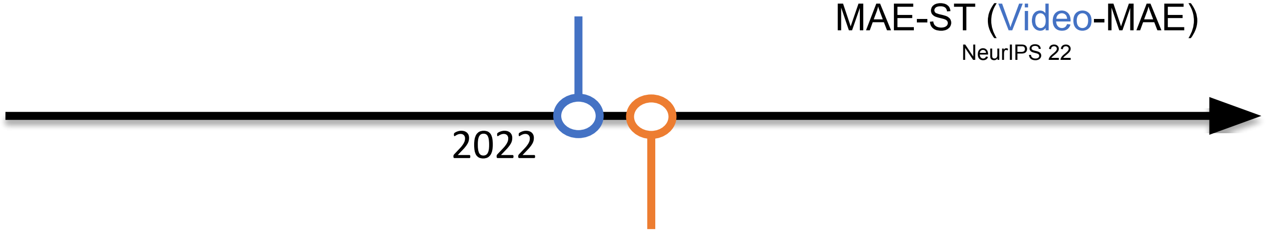


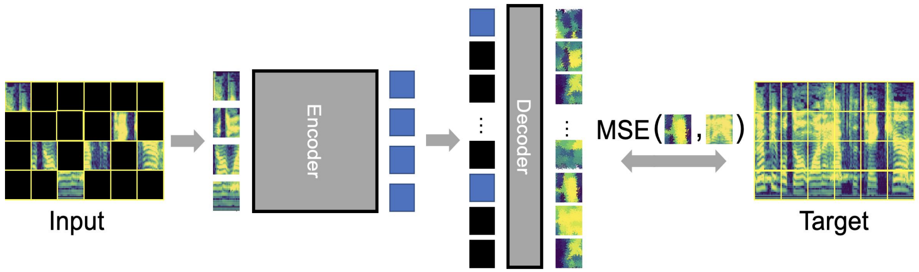
- The unifying trend
  - (data) Pre-training + Fine-tuning
  - (model) Transformer architecture
  - (learning) Large-scale *self-supervised* learning
    - SSL > SL in CV, NLP, Audio

A unifying trend across CV, NLP, Audio

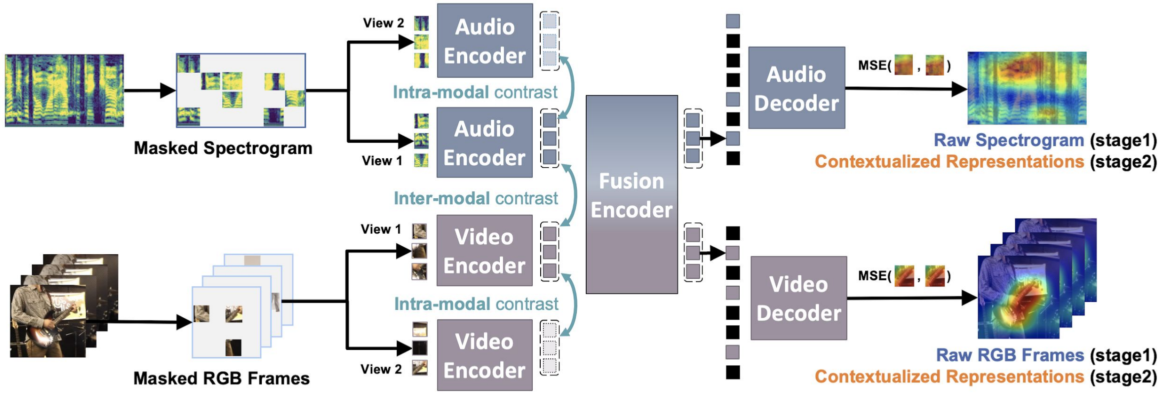


MAE-ST (Video-MAE)
NeurIPS 22





Audio-MAE
NeurIPS 22



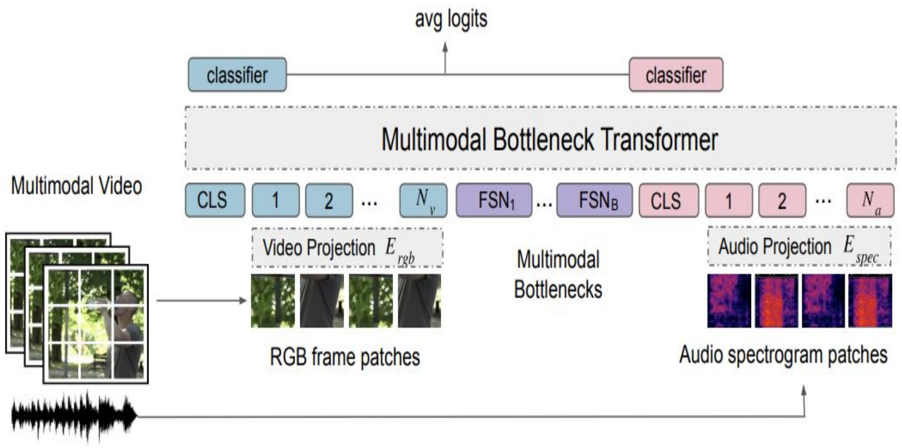
MAViL: Masked Audio-Video Learners
2022/2023

Unique challenges under the **multimodal** paradigm:

- Diverse and disparate raw inputs to MAE
- Heterogeneous context in each modality

MAViL: Masked Audio-Video Learners

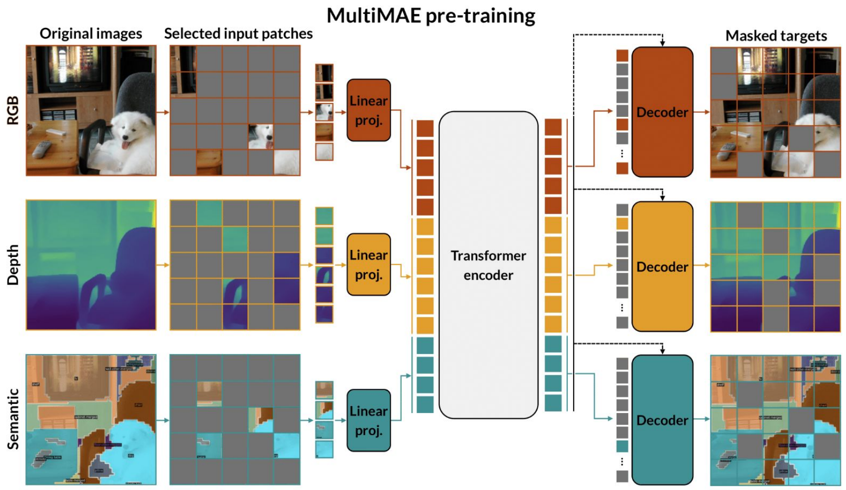
- Related work



MBT (NeurIPS 21)

Constraints:

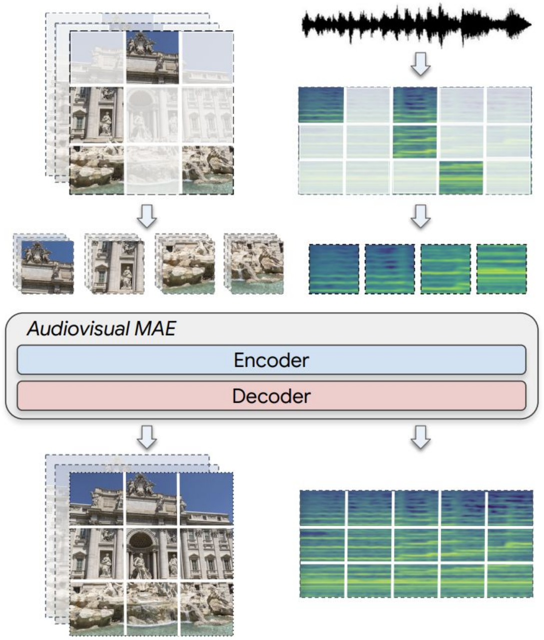
1. Supervised (labels required)
2. Computation overhead



MultiMAE (ECCV 2022)

Constraints:

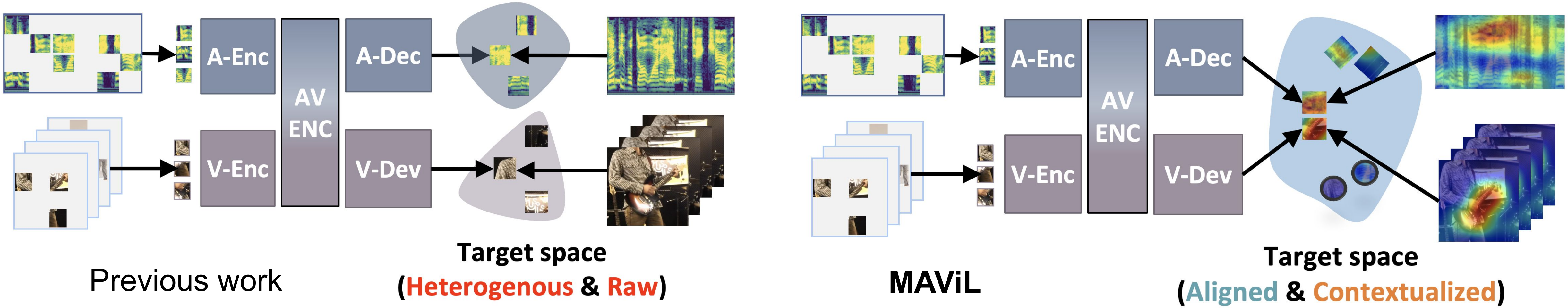
1. Naïve extension of single-modal MAE to multimodal. Limited improvement even training with multimodal context
2. Diverse and disparate raw inputs for MAE reconstruction



AudioVisual MAE (Arxiv 22)

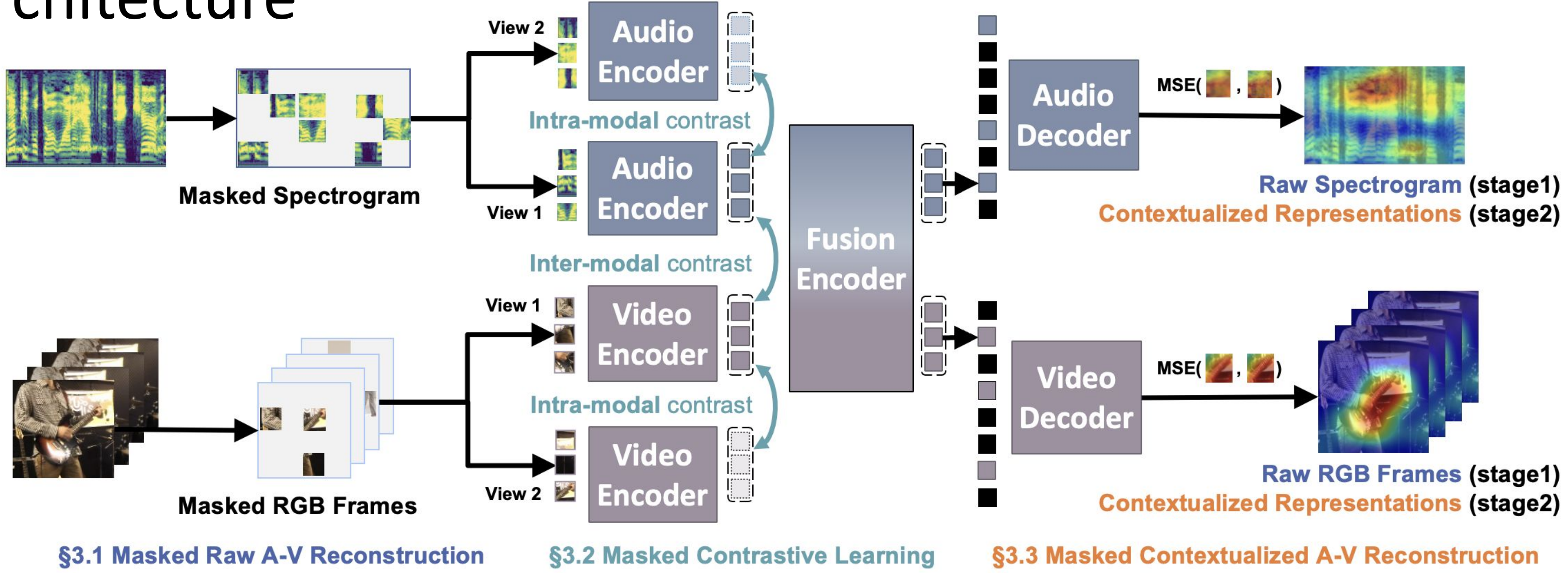
MAViL: Masked Audio-Video Learners

- Key Motivation and difference to previous multimodal-MAEs
  - ~~Reconstructing heterogenous & raw inputs~~
  - Reconstructing aligned & contextualized representations



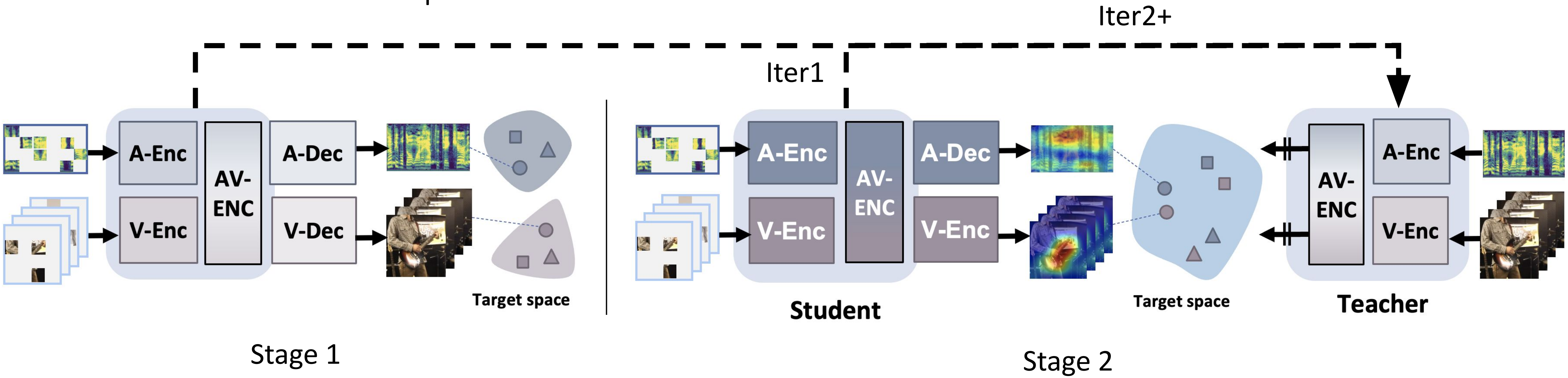
MAViL: Masked Audio-Video Learners

- Reconstructing aligned & contextualized representations
  - Inter-modal and intra-modal masked contrastive learning for promoting alignment between semantically correlated audio and/or video.
  - Train a student under masked view to predict contextualized representations in the aligned latent space generated by a teacher with full-view.
- Model Architecture



MAViL: Masked Audio-Video Learners

- Two-stage training:
  - Stage 1: Contrastive objectives and raw A-V reconstruction
  - Stage 2: Contrastive objectives and contextualized reconstruction from a Teacher
    - Iteration 1: Use Stage1 MAViL as the teacher
    - Iteration 2+: Use previous MAViL student as the teacher



Experiments

- Pre-training (PT)
  - Audioset-2M
    - 2 million 10-sec video recordings in 527 classes
      - Labels are not used (self-supervised pre-training)
    - For each 10-sec audio track
      - 128 Mel-fbanks/ 1024 time windows (stride 10 ms)
      - Shape: 1x1024x128
    - For each 10-sec video track
      - Sample 4-sec under 2fps (8 frames)
      - Shape: 8x3x224x224
  - MAViL model
    - ViT-B backbone for Audio/Video
    - 80% Masking Ratio

- Fine-tuning
  - A-V Classification
    - Audioset-20K (balanced)
    - Audioset-2M (unbalanced)
    - VGGSound
  - Audio-only Classification
    - Speech commands v1
    - ESC-50
  - Audio-Video Retrieval
    - YouCook
    - MSR-VTT

Ablation Studies

Method	Audio	Video
A-MAE/V-MAE (baseline)	36.4	17.4
MAViL stage-1		
+ Joint AV-MAE	$36.8_{(+0.4)}$	$17.7_{(+0.3)}$
+ Inter contrast	38.4	21.0
+ Intra and Inter contrast	$39.0_{(+2.2)}$	$22.2_{(+4.5)}$
MAViL stage-2		
+ Student-teacher learning	$41.8_{(+2.8)}$	$24.8_{(+2.6)}$

Observation 1: Fusing multimodal info for MAE reconstruction improve 0.3-0.4 mAP

Observation 2: Both Inter-modal and Intra-modal contrastive learning help!

Observation 3: Reconstructing aligned and contextualized representations provides additional 2.6-2.8 mAP gains!

A-V Classification

Method	PT	AS-20K (mAP↑)			AS-2M (mAP↑)			VGGSound (Acc.↑)		
		A	V	A+V	A	V	A+V	A	V	A+V
Audio-only Models										
Aud-SlowFast [68]	-	-	-	-	-	-	-	50.1	-	-
VGGSound [58]	-	-	-	-	-	-	-	48.8	-	-
PANNs [69]	-	27.8	-	-	43.9	-	-	-	-	-
AST [64]	IN-SL	34.7	-	-	45.9	-	-	-	-	-
HTS-AT [70]	IN-SL	-	-	-	47.1	-	-	-	-	-
PaSST [71]	IN-SL	-	-	-	47.1	-	-	-	-	-
Data2vec [51]	AS-SSL	34.5	-	-	-	-	-	-	-	-
SS-AST [72]	AS-SSL	31.0	-	-	-	-	-	-	-	-
MAE-AST [73]	AS-SSL	30.6	-	-	-	-	-	-	-	-
Aud-MAE [4]	AS-SSL	37.0	-	-	47.3	-	-	-	-	-
Audio-Video Models										
G-Blend [74]	-	29.1	22.1	37.8	32.4	18.8	41.8	-	-	-
Perceiver [75]	-	-	-	-	38.4	25.8	44.2	-	-	-
Attn AV [76]	IN-SL	-	-	-	38.4	25.7	44.2	-	-	-
CAV-MAE [41]	IN-SSL, AS-SSL	37.7	19.8	42.0	46.6	26.2	51.2	59.5	47.0	65.5
MBT* [27]	IN21K-SL	31.3	27.7	43.9	41.5	31.3	49.6	52.3	**51.2**	64.1
MAViL	AS-SSL	41.6	23.7	44.6	**48.7**	28.3	51.9	60.6	50.0	66.5
MAViL	IN-SSL, AS-SSL	**41.8**	**24.8**	**44.9**	**48.7**	**30.3**	**53.3**	**60.8**	50.9	**67.1**

MAViL not only learns strong joint audio-video representations (A+V), but can also improve single modality encoders *without* using the other modality during fine-tuning (A, V).

Audio-Only and Audio-Video Retrieval

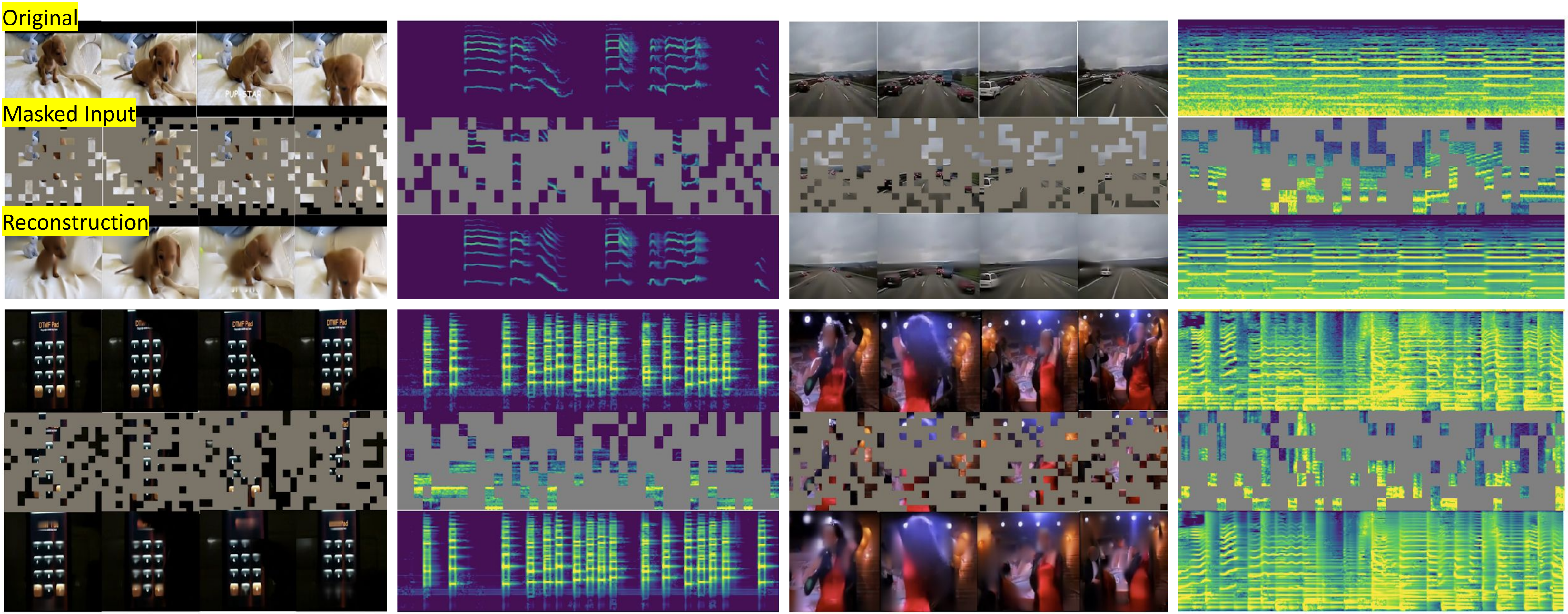
Method	PT	ESC-50	SPC-1
AST [64]	IN-SL	88.7	95.5
SS-AST [72]	AS-SSL	88.8	96.0
Aud-MAE [4]	AS-SSL	94.1	96.9
MAViL	AS-SSL	94.4	97.3
MAViL	IN-SSL, AS-SSL	**94.4**	**97.4**

Table 7: **Audio-only tasks** (Acc.↑)

Method	PT data	MSR-VTT	YouCook
AVLNet [80]	HT100M	20.1	30.7
TVLT [81]	HT100M	22.6	31.8
MAViL	AS-2M	22.8	32.2
MAViL	HT-100M	**23.8**	**33.1**

Table 8: **Audio-to-video retrieval** (R@1↑)

Qualitative Results:



MAViL Demo

Conclusion

- MAViL - MAE meets contrastive learning in Audio-Video SSL:
  - Complementary information from both audio and video are beneficial in MAE
  - Masking combined with contrastive learning demonstrates remarkable efficiency in audio-video SSL
  - Both intra-modal and inter-modal contrast learning are important
  - In the multimodal paradigm, predicting aligned and contextualized representations proves to be superior to heterogeneous and raw inputs
  - Multimodal (A+V) SSL can also improve single modality encoders *without* using the other modality during fine-tuning (A, V)
- Future Work:
  - Audio-Video-Text modeling with MAViL