ICLR 2024 Potpourri

Mert Kilickaya

ICLR 2024 received **7500** submissions on 28th September

Pre-filtering based on the title and the abstract: ~**200** papers

Post-filtering based on content: **15** papers

Four Topics: LLM & Efficiency & ViT & Industry

1) LLM-based Image Understanding

FROZEN TRANSFORMERS IN LANGUAGE MODELS ARE EFFECTIVE VISUAL ENCODER LAYERS

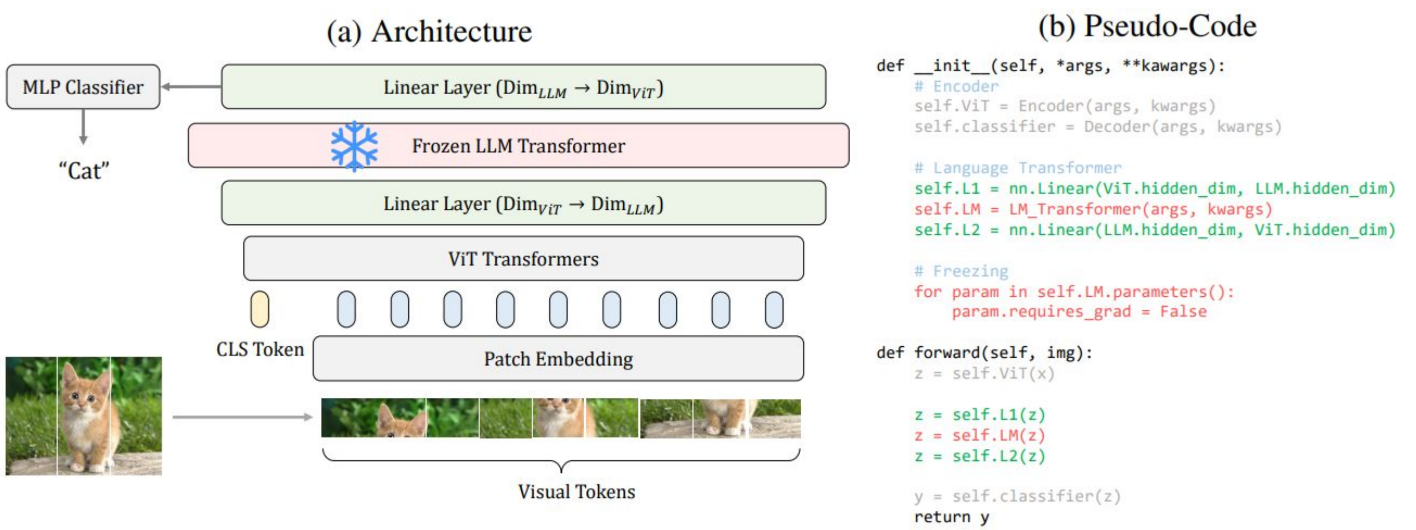


Figure 1: Our straightforward method of using a *frozen* transformer block from *pre-trained* LLMs as a visual encoder layer. Visualized with an example of ViT (Dosovitskiy et al., 2021). **(a)** Our design simply appends a frozen transformer block (pink) on top of the regular visual encoder (gray). Only two trainable linear layers (green) are added to align the feature dimensions. **(b)** Pytorch-style pseudo-code shows the simplicity of our approach.

Encode image with ViT + Linear layer **->** Process the features with (frozen) LLM **->** Classify & Detect

FROZEN TRANSFORMERS IN LANGUAGE MODELS ARE EFFECTIVE VISUAL ENCODER LAYERS

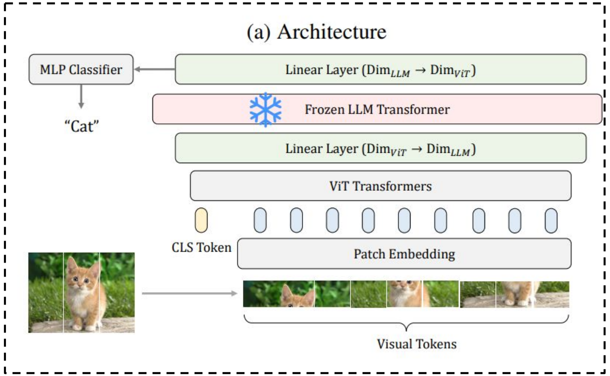
Model	ImageNet	ImageNet-C	ImageNet-A	ImageNet-SK	ImageNet-R
ViT-T	72.1	43.9	7.7	19.6	32.3
ViT-T-LLaMA	**73.2**	**45.8**	**8.7**	**20.6**	**33.8**
ViT-S	80.1	57.2	20.5	28.9	42.1
ViT-S-LLaMA	**80.7**	**58.7**	**22.7**	**30.5**	**42.8**
ViT-B*	78.9	58.1	21.6	29.3	40.5
ViT-B-LLaMA	**80.6**	**60.6**	**24.6**	**30.4**	**40.9**

The improvement is not significant (1-2%) considering the heavy burden from the LLM.

1 [Frozen-LLM]

2 [MiniGPT-V2]

3 [MiniGPT-V4]



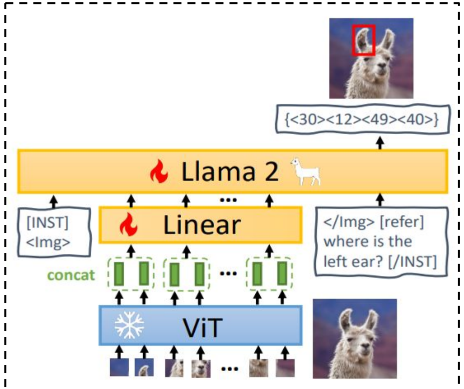
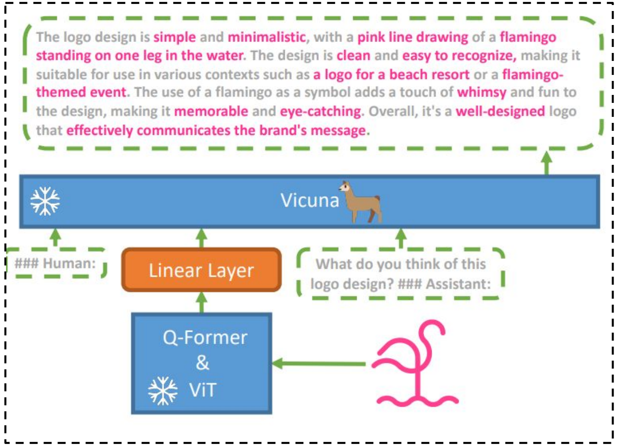


Figure 2: **Architecture of MiniGPT-v2.** The model takes a ViT visual backbone, which remains frozen during all training phases. We concatenate four adjacent visual output tokens from ViT backbone and project them into LLaMA-2 language model space via a linear projection layer.

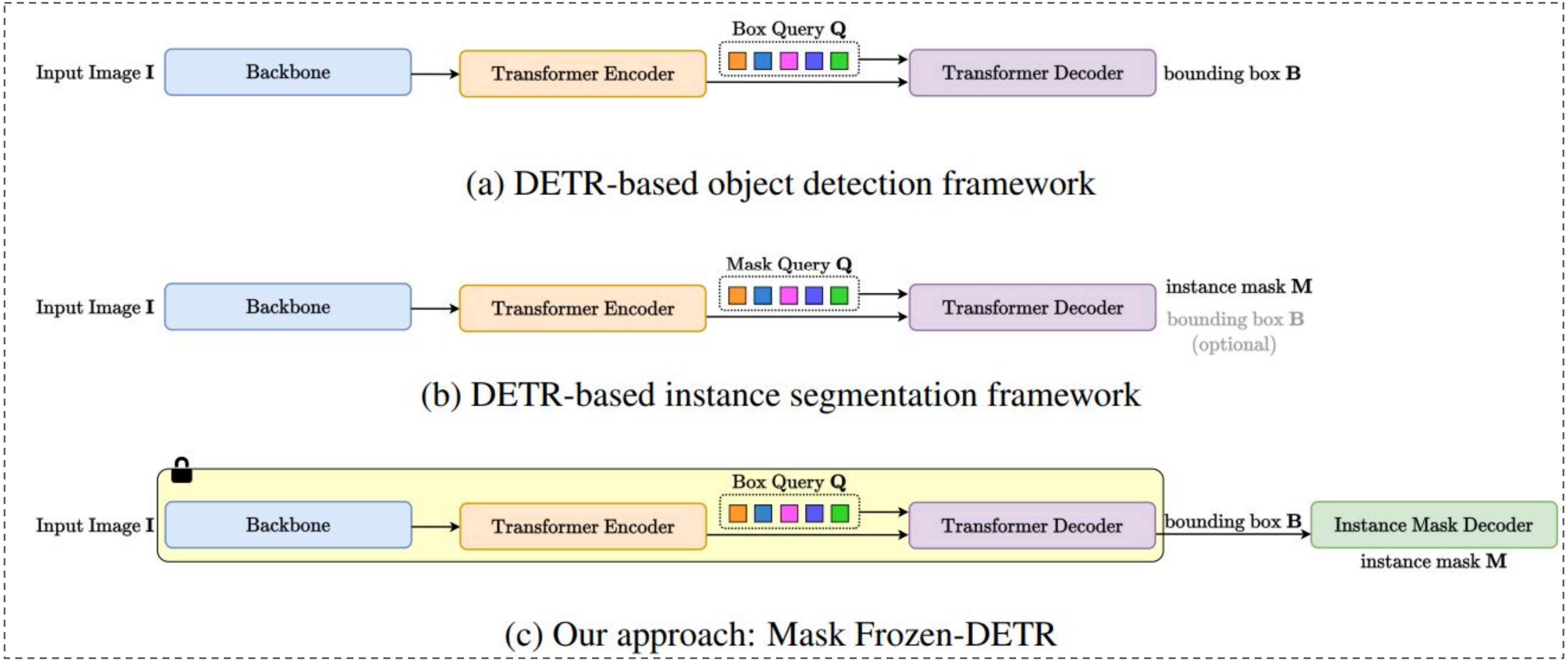


4 [PLANTING A SEED OF VISION IN LARGE LANGUAGE MODEL]

5 [LINGUISTIC IMAGE UNDERSTANDING]

2) Efficient Training & Fine-tuning

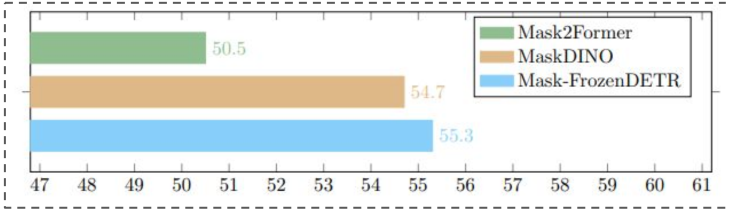
MASK FROZEN-DETR: HIGH QUALITY INSTANCE SEGMENTATION WITH ONE GPU



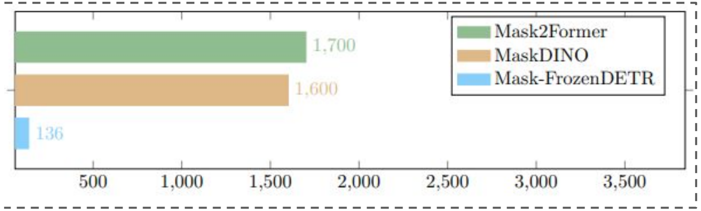
Given a pre-trained DETR detector, freeze encoder-decoder, map the output feature to segmentation mask.

MASK FROZEN-DETR: HIGH QUALITY INSTANCE SEGMENTATION WITH ONE GPU

Mask AP



GPU Hours

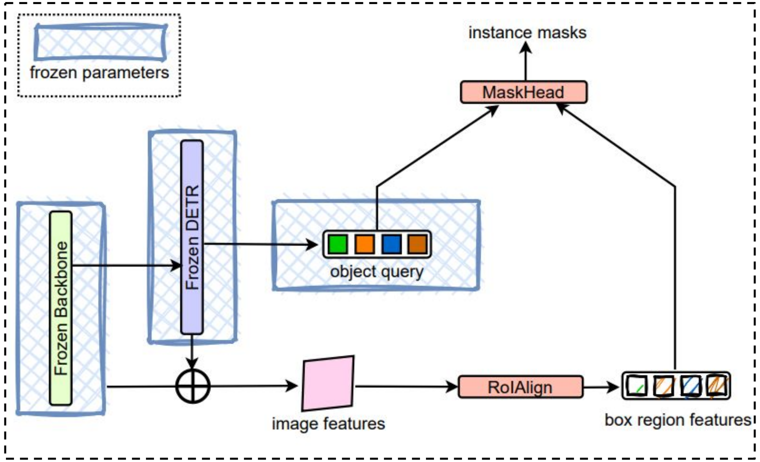


Extensive Comparison

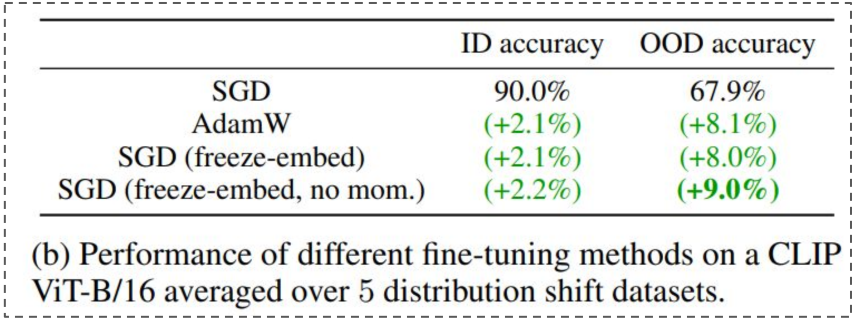
method	backbone	#epochs	Object365	AP^{box}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{mask}_{S}	AP^{mask}_{M}	AP^{mask}_{L}	GPU Hours
K-Net-N256 (Zhang et al., 2021)	R50	36	✗	–	38.6	60.9	41.0	19.1	42.0	57.7	–
QueryInst (Fang et al., 2021)	Swin-L	50	✗	56.1	48.9	74.0	53.9	30.8	52.6	68.3	–
Mask2Former (Cheng et al., 2021a)	R50	50	✗	–	43.7	–	–	23.4	47.2	64.8	502
Mask2Former (Cheng et al., 2021a)	Swin-L	100	✗	–	50.1	–	–	29.9	53.9	72.1	1, 700
Mask DINO (Li et al., 2022)	R50	50	✗	50.5	46.0	68.9	50.3	26.0	49.3	65.5	1, 404
Mask DINO (Li et al., 2022)	Swin-L	50	✗	58.3	52.1	76.5	57.6	32.9	55.4	72.5	2, 400
Mask Frozen-$\mathcal{H}$-DETR	R50	6	✗	49.9	44.1	66.2	47.8	24.4	47.0	62.6	49
Mask Frozen-$\mathcal{H}$-DETR	Swin-L	6	✗	59.1	51.9	75.8	57.2	31.6	55.1	71.6	179
ViT-Adapter-L (Chen et al., 2022)	ViT-L	8	✓	61.8	53.0	–	–	–	–	–	1, 068
Mask DINO (Li et al., 2022)	Swin-L	24	✓	–	54.5	–	–	–	–	–	1, 600
Mask Frozen-$\mathcal{H}$-DETR	R50	6	✓	52.2	45.7	67.5	49.8	25.6	48.9	64.1	49
Mask Frozen-$\mathcal{H}$-DETR	Swin-L	6	✓	62.3	54.0	77.9	59.5	35.6	57.4	73.0	172
Mask Frozen-DINO-DETR	FocalNet-L	6	✓	63.2	**54.9**	**78.9**	**60.8**	**37.2**	**58.4**	**72.9**	136

MaskFrozen outperforms MaskDINO, while training much more efficiently (i.e. 10x less GPU hours)

1 Mask-Frozen DETR



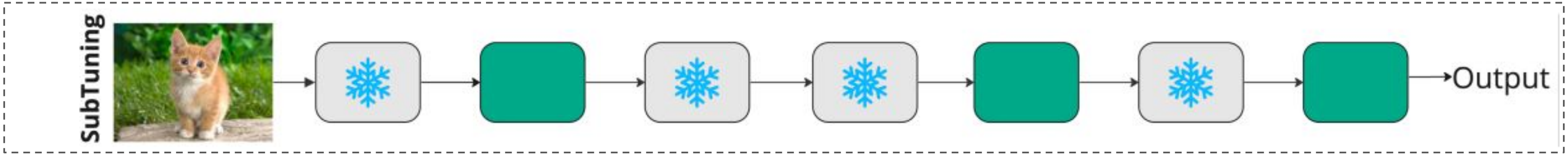
2 HOW TO FINE-TUNE VISION MODELS WITH SGD



	ID accuracy	OOD accuracy
SGD	90.0%	67.9%
AdamW	(+2.1%)	(+8.1%)
SGD (freeze-embed)	(+2.1%)	(+8.0%)
SGD (freeze-embed, no mom.)	(+2.2%)	**(+9.0%)**

(b) Performance of different fine-tuning methods on a CLIP ViT-B/16 averaged over 5 distribution shift datasets.

3 LESS IS MORE: SELECTIVE LAYER FINETUNING WITH SUBTUNING



3) Vision-Transformers

FASTERVIT: FAST VISION TRANSFORMERS WITH HIERARCHICAL ATTENTION

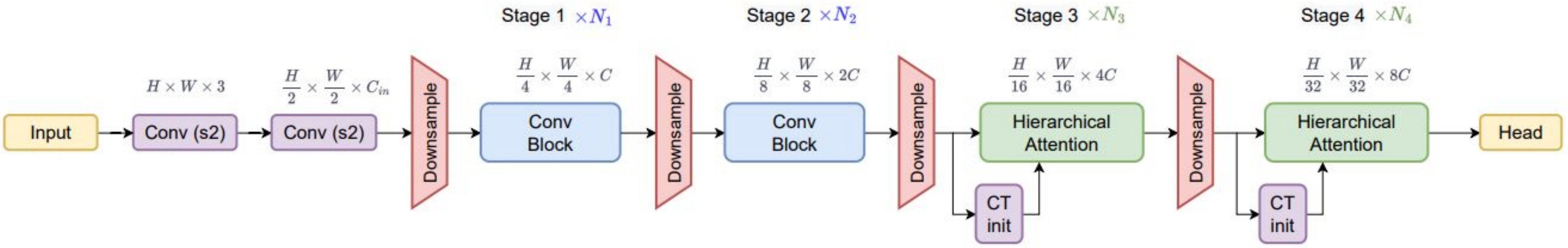
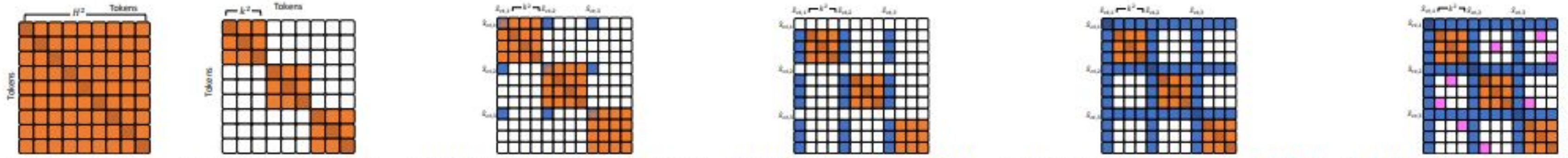


Figure 3: Overview of the FasterViT architecture. We use a multi-scale architecture with CNN and transformer-based blocks in stages 1, 2 and 3, 4, respectively. Best viewed in color.

CT Init: Carrier token initialization

FasterViT combines Convolutional blocks (stage 1-2) with Transformer blocks (stage 3-4).

FASTERVIT: FAST VISION TRANSFORMERS WITH HIERARCHICAL ATTENTION



(a) Full attention (b) Windowed attention (c) **Hierarchical Attention (ours)** (d) Twins Chu et al. (2021a) (e) LongFormer Beltagy et al. (2020) (f) BigBird Zaheer et al. (2020)

Figure 5: Attention map comparison for a feature map of size $H \times H \times d$. □ - no attention, ■ - normal token attention, ■ - carrier token attention, ■ - random token attention. Full attention (a) has complexity of $O(H^4 d)$, windowed attention significantly reduces it to $O(k^2 H^2 d)$ but lacks global context.

Hierarchical attention extends windowed attention via carrie tokens to pool from distinct windows.

FASTERVIT: FAST VISION TRANSFORMERS WITH HIERARCHICAL ATTENTION

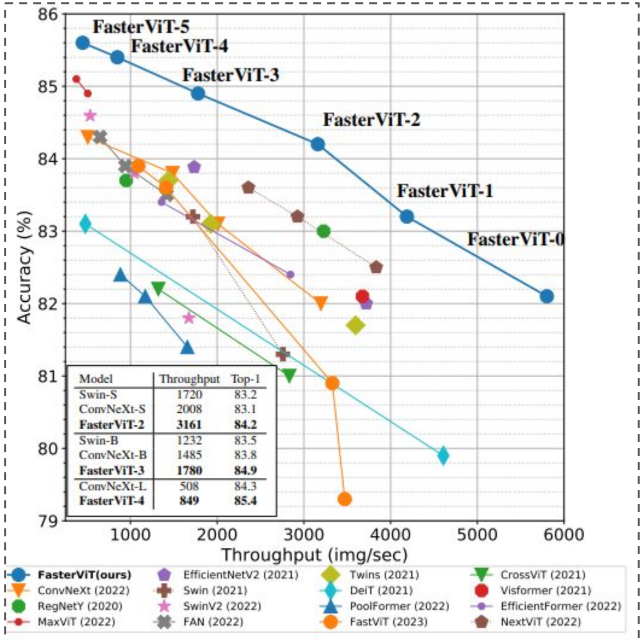


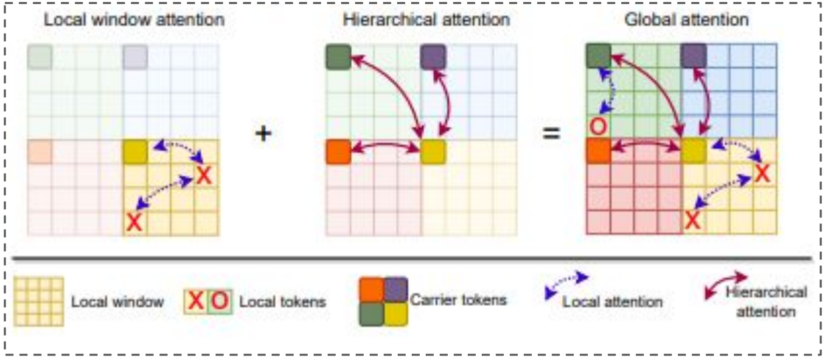
Figure 1: Comparison of image throughput and ImageNet-1K Top-1 accuracy. For all models, throughput is measured on A100 GPU with batch size of 128.

Model	Throughput	FLOPs (G)	IoU(ss/ms)
Swin-T Liu et al. (2021)	350	945	44.5/45.8
ConvNeXt-T Liu et al. (2022b)	363	939	- /46.7
FasterViT-2	**377**	974	**47.2/48.4**
Twins-SVT-B Chu et al. (2021a)	204	-	47.7/48.9
Swin-S Liu et al. (2021)	219	1038	47.6/49.5
ConvNeXt-S Liu et al. (2022b)	234	1027	- /49.6
FasterViT-3	**254**	1076	**48.7/49.7**
Twins-SVT-L Chu et al. (2021a)	164	-	48.8/50.2
Swin-B Liu et al. (2021)	172	1188	48.1/49.7
ConvNeXt-B Liu et al. (2022b)	189	1170	- /49.9
FasterViT-4	**202**	1290	**49.1/50.3**

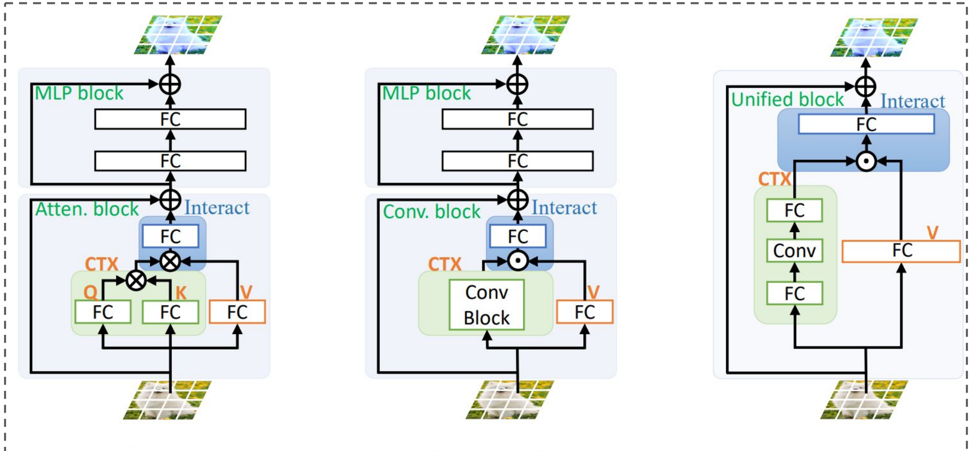
Table 4: Semantic segmentation on **ADE20K** Zhou et al. (2017) with UPerNet Xiao et al. (2018).

FasterViT yields the highest accuracy-efficiency tradeoff for classification/segmentation tasks (ADE20K).

1 [FASTERVIT]



2 [EFFICIENT MODULATION FOR VISION NETWORKS]



(a) Transformer block (b) Modulation design (c) Our EfficientMod block

3 [CHANNEL VISION TRANSFORMERS: AN IMAGE IS WORTH C × 16 × 16 WORDS]

4) Industrial Vision

DEFECT SPECTRUM: A GRANULAR LOOK OF LARGE-SCALE DEFECT DATASETS WITH RICH SEMANTICS

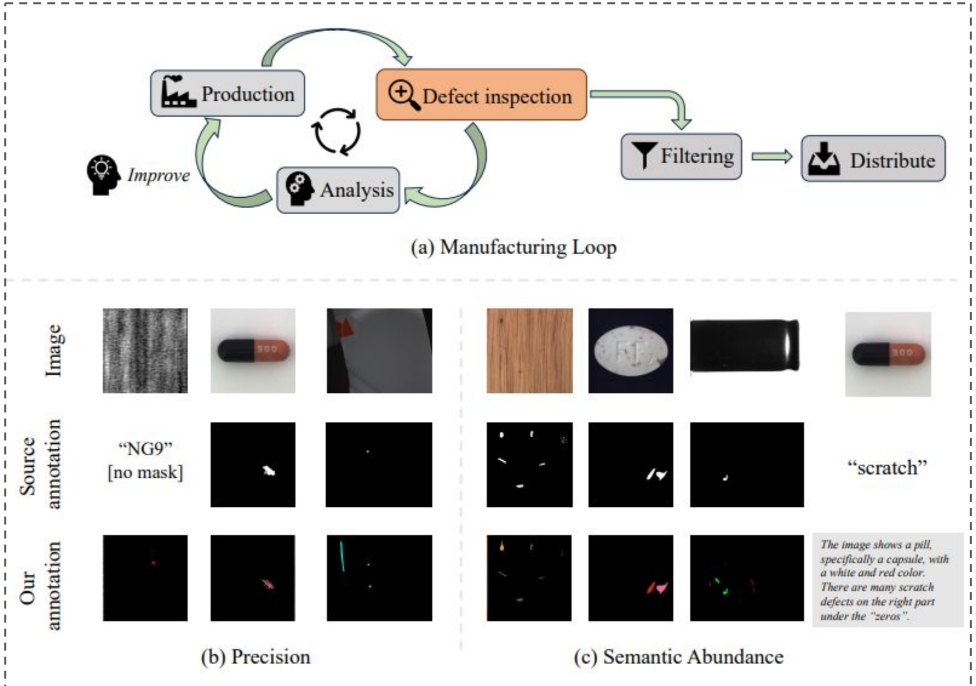


Figure 1: (a) The closed-loop system in industrial manufacturing. Defect inspection plays a pivotal role. (b, c) An overview of our improvements in annotations, in the aspect of precision and semantics abundance. **Best viewed in color.**

The authors present: 1) Fine-grained defect annotations, 2) Synthetic defect generator via diffusion models.

DEFECT SPECTRUM: A GRANULAR LOOK OF LARGE-SCALE DEFECT DATASETS WITH RICH SEMANTICS

Dataset Annotation Comparison

	Annotated Defective Images	Defect Type	Pixel-wise Label	Multiple Defective Label	Detailed Caption
AITEX	105	12	✓		
AeBAD	346	4	✓		
BeanTech	290	3	✓		
Cotton-Fabric	89	1			
DAGM2007	900	6			
KolektorSDD2	356	1	✓		
MVTec	1258	69	✓		
VISION V1	4165	44	✓	✓	
Defect Spectrum	3518+1920*	125	✓	✓	✓

Training w/ synthetic data improves performance

Table 4: Performance (mIoU) comparison between models trained with and without synthetic data.

	MVTec	VISION	Cotton
w/o synthetic	51.58	54.12	64.09
w. synthetic	55.55	55.47	65.39

Comparison of inference time across models. **HR**Netw18small is the fastest.

Table 2: Speed Evaluation of the baselines. Inf. time denotes the inference time of a single image on NVIDIA RTX 3090.

	UNet	PSP	DL	HR	Bise	V-T	M-B0	M2F
inf. time (ms)	33.9	26.2	33.0	15.7	23.5	38.7	17.9	68.2

SLIME: SEGMENT LIKE ME

Figure 1: ***SLiMe*.** Using just one user-annotated image with various granularity (as shown in the leftmost column), *SLiMe* learns to segment different unseen images in accordance with the same granularity (as depicted in the other columns).

SLIME can learn to segment objects at different granularity from a single image, interactively.

1 [DEFECT SPECTRUM]

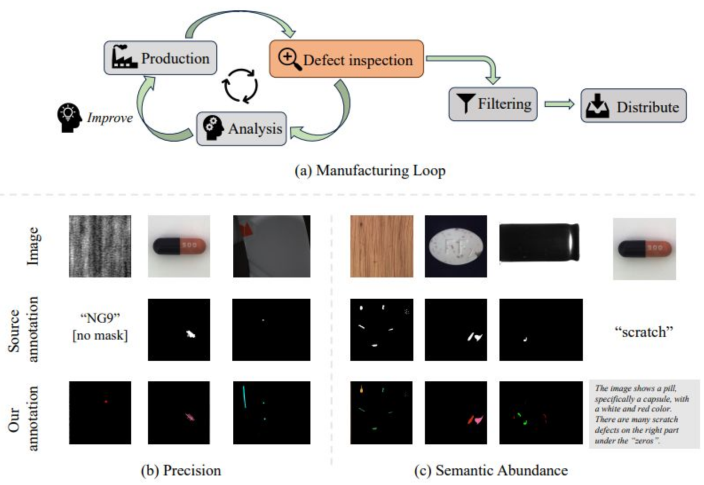


Figure 1: (a) The closed-loop system in industrial manufacturing. Defect inspection plays a pivotal role. (b, c) An overview of our improvements in annotations, in the aspect of precision and semantics abundance. **Best viewed in color.**

2 [SLIME: SEGMENT LIKE ME]

Figure 1: ***SLiMe.*** Using just one user-annotated image with various granularity (as shown in the leftmost column), *SLiMe* learns to segment different unseen images in accordance with the same granularity (as depicted in the other columns).

3 [ANOMALYCLIP: OBJECT-AGNOSTIC PROMPT LEARNING FOR ZERO-SHOT ANOMALY DETECTION]

4 [YOLOV6: A SINGLE-STAGE OBJECT DETECTION FRAMEWORK FOR INDUSTRIAL APPLICATIONS]

Thanks! Questions?