MAE-DET: Revisiting Maximum Entropy Principle in Zero-Shot NAS for Efficient Object Detection

Zhenhong Sun [*1] **Ming Lin** [*1] **Xiuyu Sun** [1] **Zhiyu Tan** [1] **Hao Li** [1] **Rong Jin** [1]

Contact us by DingTalk

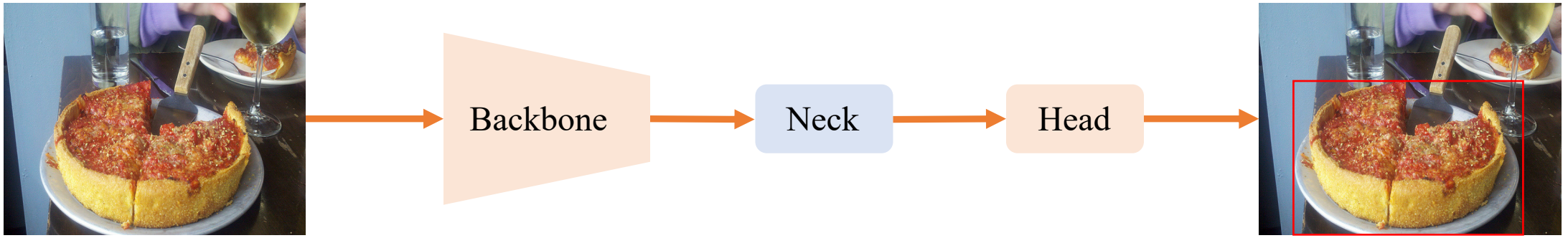
Contact us by WeChat

Zhenhong Sun
Algorithm Engineer
DAMO Academy, Alibaba Group

Outline

- ✓Motivation
- ✓Maximum Entropy Principle
- ✓Single-scale Entropy for Deep Networks
- ✓Multi-scale Entropy for Object Detection
- ✓Evolutionary Algorithm for MAE-DET
- ✓Experimental Results
- ✓Conclusion

Motivation -- Backbone of Object Detection



- ✓The performance of object detection network heavily depends on the feature extraction backbone.
- ✓SOTA detection backbones are designed manually by human experts, migrated from classification.
- ✓Since backbone consumes more than half of the overall inference cost, it is critical to optimize the backbone for better speed-accuracy trade-off on different hardware platforms.

Motivation -- Two Challenges for Object Detection

- ✓ Maximum Entropy Principle can indicate the expressivity of a network.
  - Regard a detection network as an information processing system, its expressivity is maximized when its entropy achieves maximum under the given inference budgets.
  - The maximum expressivity represents a better feature extractor for object detection.
- ✓ Two challenges to apply the entropy to Traning-free detection NAS.
  - How to estimate the entropy of a deep network?
  - How to efficiently extract deep features for objects of different scales?

Maximum Entropy Principle -- Expressivity

Alibaba Group 阿里巴巴集团

✓ Continuous State Space of Deep Networks

- Deep network $F(\cdot): \mathbb{R}^{\mathrm{d}} \rightarrow \mathbb{R}$ maps an input image $\boldsymbol{x} \in \mathbb{R}^{\mathrm{d}}$ to its label $\boldsymbol{y} \in \mathbb{R}$.
- $S = \{h(v), h(e): \forall v \in \mathcal{V}, e \in \mathcal{E}\}$ defines the continuous state space of the network F .
- $H(S_v)$ measures the feature representation power, representing the expressivity.
- $H(S_e)$ measures the network parameters, representing model complexity.

✓ Gaussian Entropy Upper Bound

- *Theorem*: For any continuous distribution $\mathbb{P}(x)$ of mean μ and variance σ^2, its differential entropy is maximized when $\mathbb{P}(x)$ is a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$.

Maximum Entropy Principle -- Expressivity

Alibaba Group 阿里巴巴集团

✓ Entropy of Gaussian Distribution

- Suppose x is sampled from Gaussian distribution N (μ, σ2). Then the differential entropy of x is given by

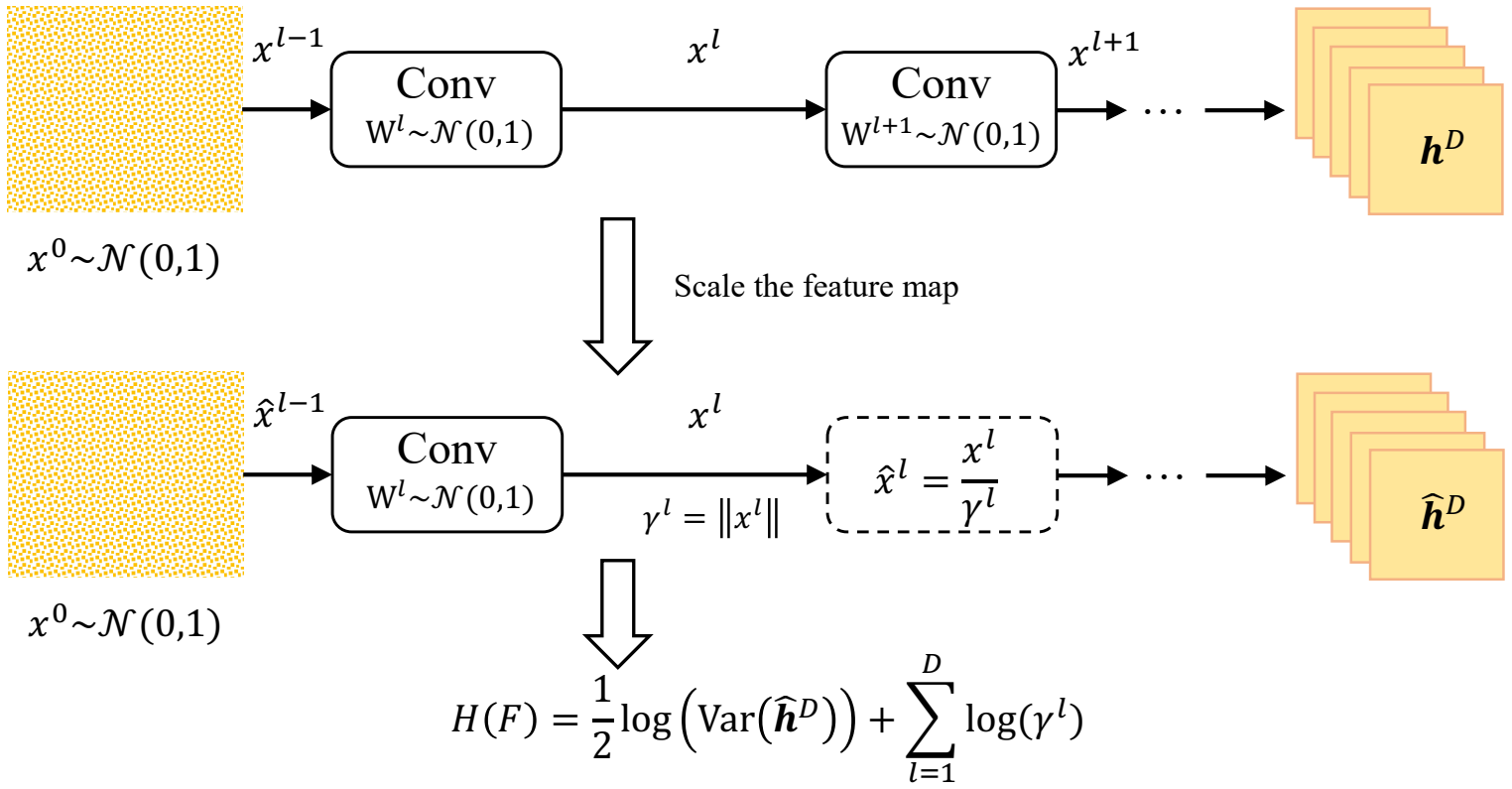
$$H^*(x) = \frac{1}{2}\log(2\pi) + \frac{1}{2} + H(x) \qquad H(x) \coloneqq \log(\sigma).$$

✓ Vanilla Network Search Space

- A vanilla network is stacked by multiple convolutional layers, followed by RELU activations with bias set to zero:

$$x^l = \varphi(h^l), \qquad h^l = W^l * x^{l-1}, l = 1, \dots, D.$$

Single-scale Entropy for Deep Networks



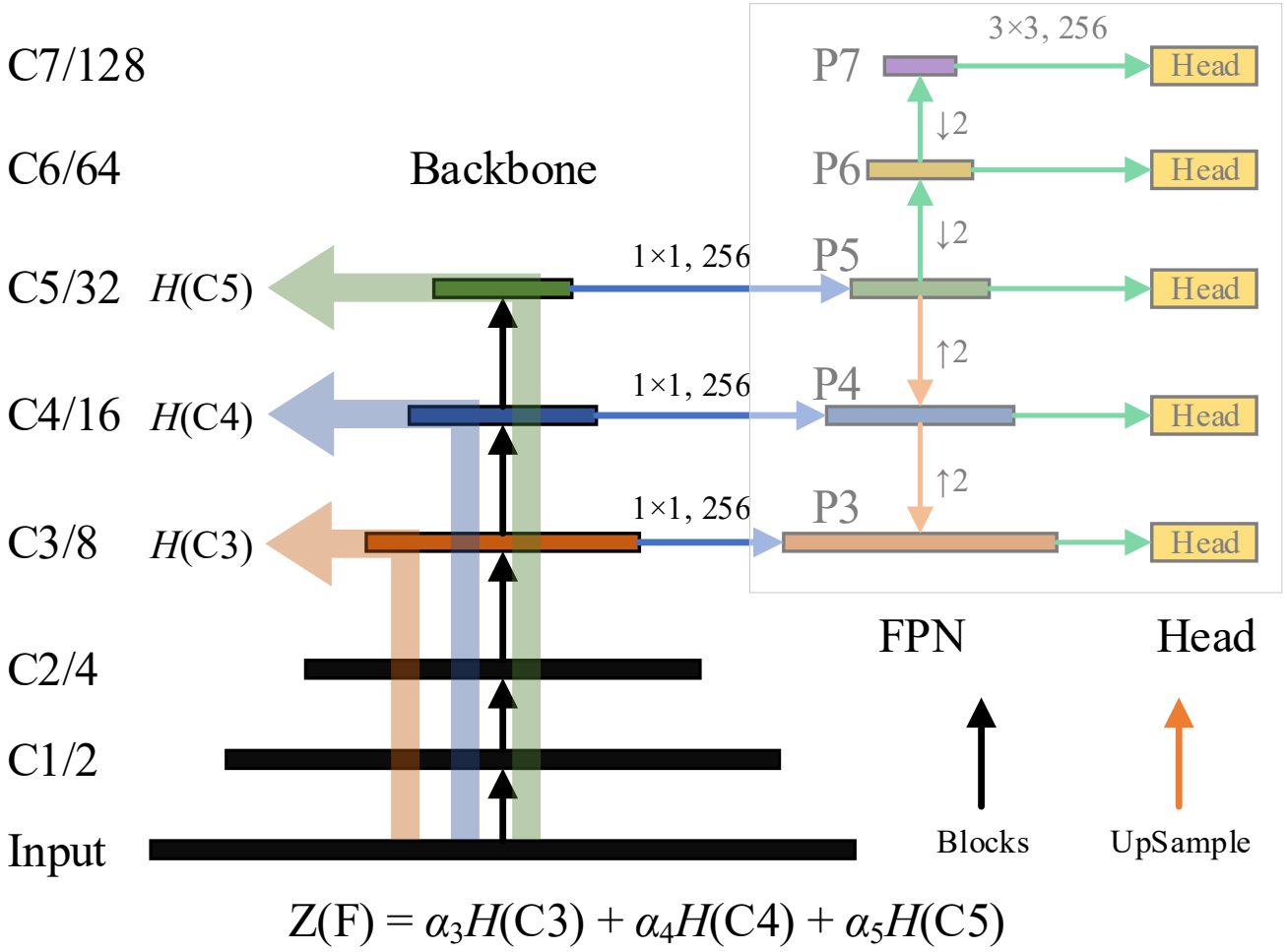
$$H(F) = \frac{1}{2}\log(\mathrm{Var}(\boldsymbol{h}^D)) .$$

Scaling

$$H(F) = \frac{1}{2}\log(\mathrm{Var}(\hat{\boldsymbol{h}}^D)) + \sum_{l=1}^{D}\log(\gamma^l) .$$

- ✓ Parameters are initialized by the standard Gaussian distribution.
- ✓ Randomly generate an image input filled with the standard Gaussian noise.
- ✓ Perform forward inference to calculate the Gaussian upper bound entropy of the network.

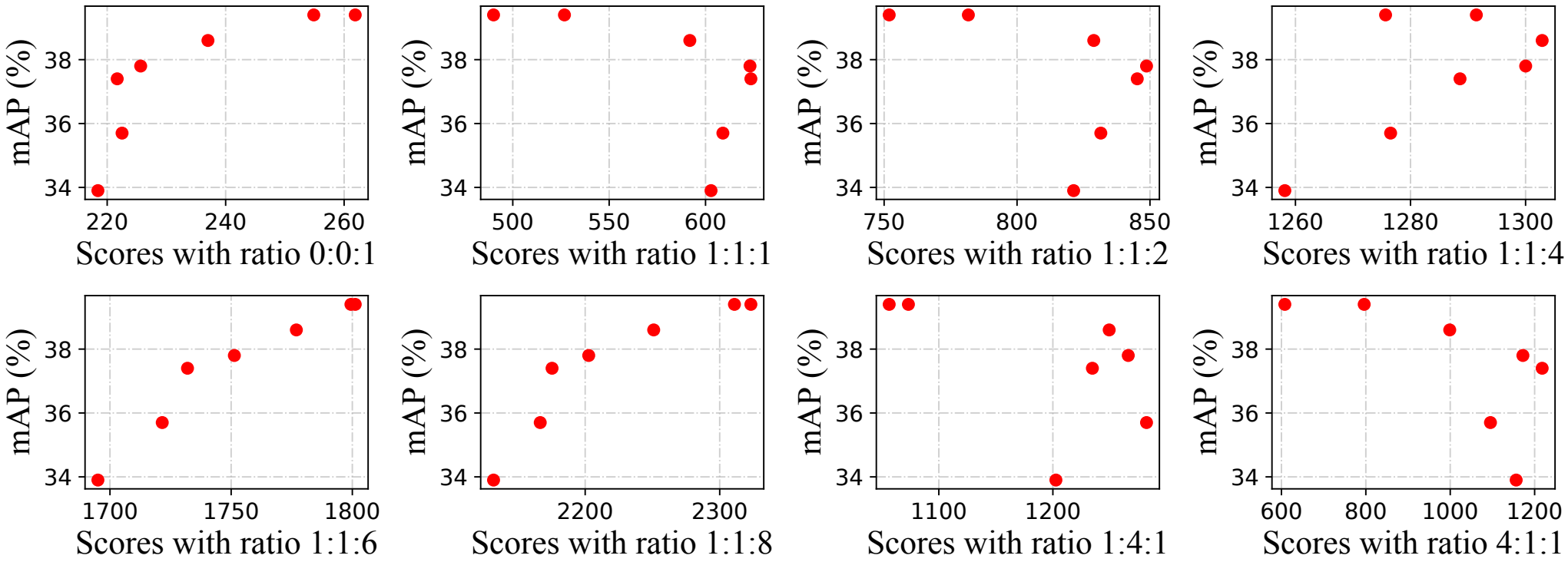
Multi-scale Entropy for Object Detection



$$Z(F) = \alpha_3 H(C3) + \alpha_4 H(C4) + \alpha_5 H(C5)$$

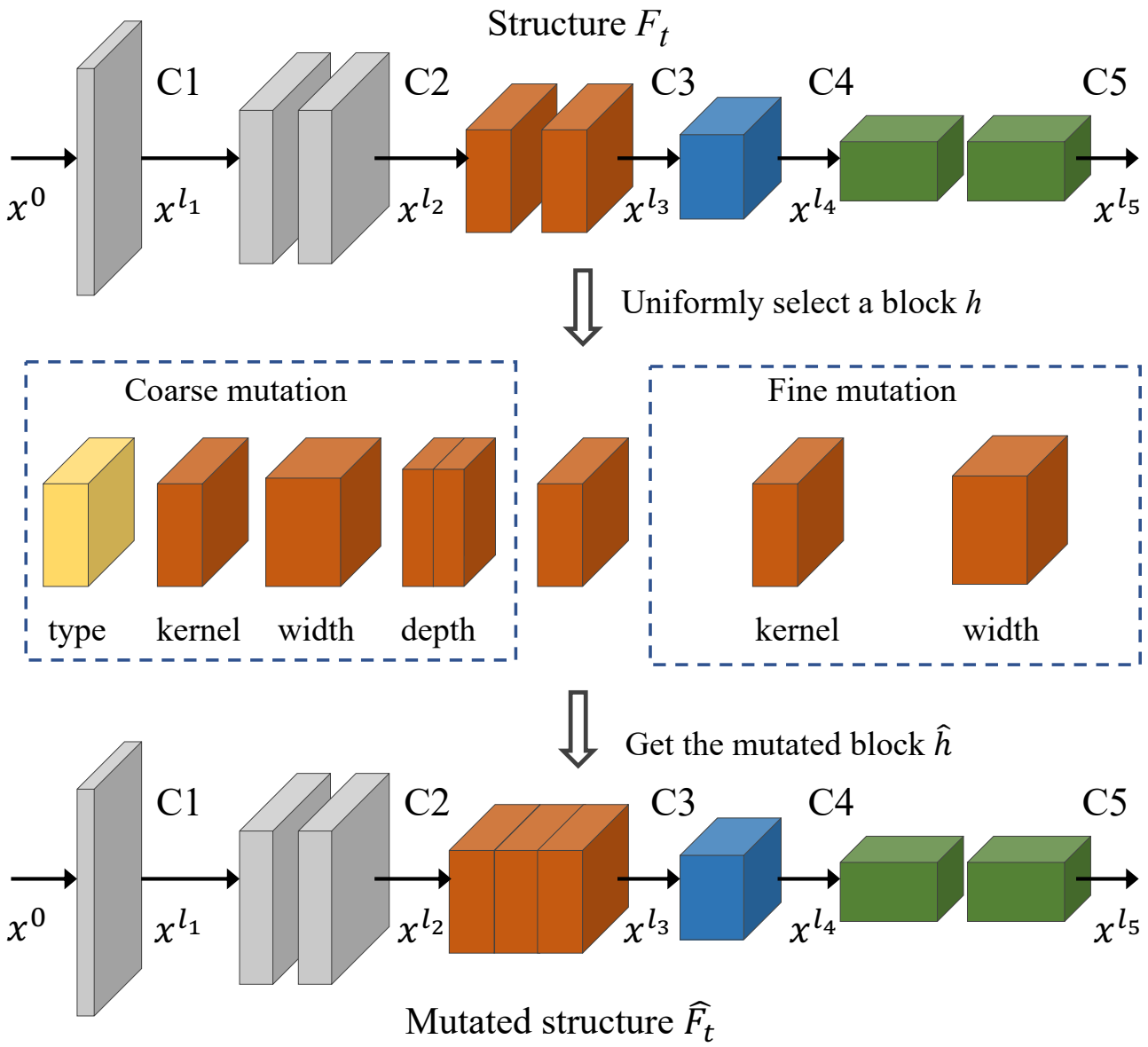
- ✓Multi-scale features C at different resolutions for Detection.
- ✓FPN neck fuses C into cross-stage features P to exchange the Info.
- ✓C5 is more important (up and down).
- ✓Weights α store the multi-scale entropy prior to balance the expressivity.

Multi-scale Entropy for Object Detection



- ✓ Explore different combinations of α and correlation analysis.
- ✓ α = (0, 0, 1, 1, 6) is good enough for the FPN structure.

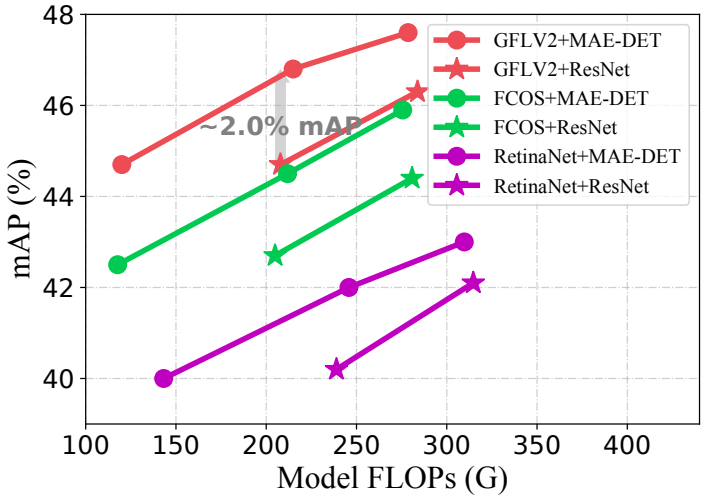
Evolutionary Algorithm for MAE-DET

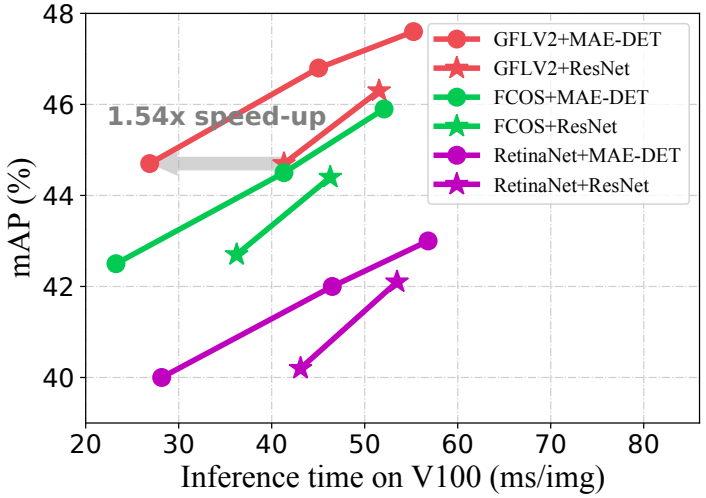


- ✓ Evolutionary Algorithm
  - Small initial network
  - Stacked ResNet or MBV2 blocks
  - Fixed maximum length
  - Coarse2Fine mutation
  - Maintain the population

Experimental Results -- Better ResNet-like Backbones

Backbone	FLOPs Backbone	Params Backbone	Head	val2017 AP_{val}	AP_S	AP_M	AP_L	test-dev AP_{test}	FPS on V100
ResNet-50	83.6G	23.5M	RetinaNet	40.2	24.3	43.3	52.2	-	23.2
			FCOS	42.7	28.8	46.2	53.8	-	27.6
			GFLV2	44.7	29.1	48.1	56.6	45.1	24.2
ResNet-101	159.5G	42.4M	RetinaNet	42.1	25.8	45.7	54.1	-	18.7
			FCOS	44.4	28.3	47.9	56.9	-	21.6
			GFLV2	46.3	29.9	50.1	58.7	46.5	19.4
MAE-DET-S	48.7G	21.2M	RetinaNet	40.0	23.9	43.3	52.7	-	35.5
			FCOS	42.5	26.8	46.0	54.6	-	43.0
			GFLV2	44.7	27.6	48.4	58.2	44.8	37.2
MAE-DET-M	89.9G	25.8M	RetinaNet	42.0	26.7	45.2	55.1	-	21.5
			FCOS	44.5	28.6	48.1	56.1	-	24.2
			GFLV2	46.8	29.9	50.4	60.0	46.7	22.2
MAE-DET-L	152.9G	43.9M	RetinaNet	43.0	27.3	46.5	56.0	-	17.6
			FCOS	45.9	30.2	49.4	58.4	-	19.2
			GFLV2	47.6	30.2	51.8	60.8	48.0	18.1





✓ Better than ResNet-series backbone under three common frameworks.

Experimental Results -- Ablation Study

Table 4. Comparison of different evolutionary searching strategies in MAE-DET. C-to-F: Coarse-to-Fine. Zen-Score is the proxy in Zen-NAS (Lin et al., 2021).

Score	Mutation	ImageNet-1K			COCO with YOLOF				COCO with FCOS			
		FLOPs	Params	TOP-1 %	AP_{val}	AP_S	AP_M	AP_L	AP_{val}	AP_S	AP_M	AP_L
ResNet-50	None	4.1G	23.5M	78.0	37.8	19.1	42.1	53.3	38.0	23.2	40.8	47.6
Zen-Score	Coarse	4.4G	67.9M	78.9	38.9	19.0	43.2	56.0	38.1	23.2	40.5	48.1
Single-scale	Coarse	4.4G	**60.1M**	78.7	39.8	19.9	44.4	56.5	38.8	23.1	41.4	50.1
Multi-scale	Coarse	4.3G	29.4M	78.9	40.1	21.1	44.5	55.9	39.4	23.7	42.3	50.0
Multi-scale	C-to-F	4.4G	25.8M	**79.1**	**40.3**	**20.8**	**44.7**	**56.4**	**40.0**	**24.5**	**42.6**	**50.6**

- ✓ Single-scale model has better performance than ResNet-50 on ImageNet, YOLOF and FCOS.
- ✓ Single-scale model has better performance than Zen-score on YOLOF and FCOS.
- ✓ Multi-scale with C-to-F strategy get the best performance on all tasks.

Experimental Results -- SOTA NAS Methods

Table 2. Comparisons with SOTA NAS methods for object detection. FLOPs are counted for full detector.

Method	Training-free	Search Cost GPU Days	Search Part	FLOPs All	Pretrain/ Scratch	Epochs	COCO ($\mathbf{AP}_{\mathbf{test}}$)
DetNAS	×	68	backbone	289G	Pretrain	24	43.4
SP-NAS	×	26	backbone	655G	Pretrain	24	47.4
SpineNet	×	100x TPUv3†	backbone+FPN	524G	Scratch	350	48.1
MAE-DET	✓	**0.6**	backbone	279G	Scratch	73	48.0

Table 3. Comparisons between MAE-DET, DetNAS (Chen et al., 2019b) and SpineNet (Du et al., 2020) under the same training settings. All backbones are trained under GFLV2 head with 6X training epochs. FLOPs and parameters are counted for full detector.

Backbone	Search Part	Search Space	FLOPs	Params	$\mathbf{AP}_{\mathbf{val}}$	$\mathbf{AP}_S$	$\mathbf{AP}_M$	$\mathbf{AP}_L$	FPS on V100
DetNAS-3.8G	backbone	ShuffleNetV2 +Xception	205G	35.5M	46.4	29.3	50.0	59.0	17.6
SpineNet-96	backbone+FPN	ResNet Block	216G	41.3M	46.6	29.8	50.2	58.9	19.9
MAE-DET-M	backbone	ResNet Block	215G	34.9M	46.8	29.9	50.4	60.0	22.2

- ✓ MAE-DET achieves better mAP than DetNAS and SP-NAS while being 50 ~ 100 times faster in search.
- ✓ MAE-DET requires fewer parameters and has a faster inference speed on V100 when achieving competitive performance over DetNAS and SpineNet on COCO.

Experimental Results -- Transfer to Other Tasks

Table 5. Transferability of MAE-DET in multiple object detection and instance segmentation tasks. FLOPs reported are counted for full detector.

Task	Dataset	Head	Backbone	Resolution	Epochs	FLOPs	AP_{val}	AP^{mask}_{val}
Object Detection	VOC	FCOS	ResNet-50	1000 × 600	12	120G	76.8	-
			MAE-DET-M	1000 × 600	12	123G	**80.9**	-
	Citescapes		ResNet-50	2048 × 1024	64	411G	37.0	-
			MAE-DET-M	2048 × 1024	64	426G	**38.1**	-
Instance Segmentation	COCO	MASK R-CNN	ResNet-50	1333 × 800	73	375G	43.2	39.2
			MAE-DET-M	1333 × 800	73	379G	**44.5**	**40.3**
			ResNet-50†	640 × 640	350	228G	42.7	37.8
			SpineNet-49†	640 × 640	350	216G	42.9	38.1
		SCNet	ResNet-50	1333 × 800	73	671G	46.3	41.6
			MAE-DET-M	1333 × 800	73	675G	**47.1**	**42.3**

†: Numbers are cited from SpineNet paper (Du et al., 2020).

✓ While transferring to VOC and Cityscapes dataset, MAE-DET achieves better performanc than ResNet-50.

✓ While transferring to COCO instance segmentation, MAE-DET still works well.

Conclusion

- ✓ We revisit the Maximum Entropy Principle in zero-shot object detection NAS, and deliver superior performance without bells and whistles.
- ✓ Using less than one GPU day and 2GB memory, MAE-DET achieves competitive performance on COCO with at least 50x times faster.
- ✓ MAE-DET is the first zero-shot NAS method for object detection with SOTA performance under multiple detection frameworks.

Alibaba Group 阿里巴巴集团

Contact us by DingTalk

Contact us by WeChat

Thank you for your listening