MetaUAS: Universal Anomaly Segmentation with One-Prompt Meta-Learning

Bin-Bin Gao, Tencent YouTu Lab, csgaobb@gmail.com

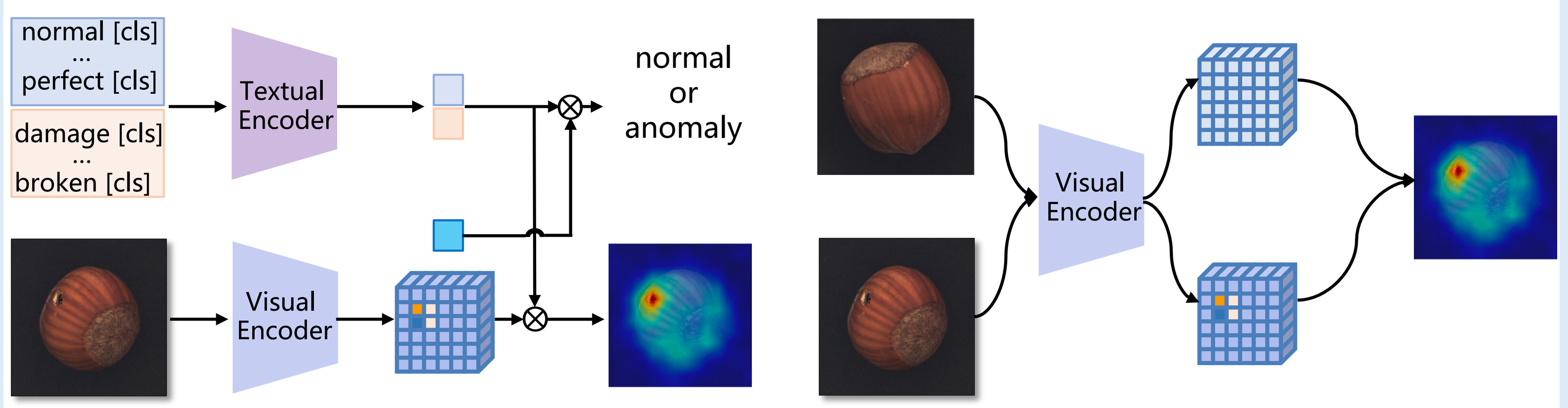
Tencent YouTu Lab

MetaUAS Code

NEURAL INFORMATION PROCESSING SYSTEMS

1. Introduction

Zero-/few-shot anomaly segmentation aims to identify any novel anomalies within zero or only a few normal images. Most methods relies on powerful vision-language models using manually designed textual prompts.



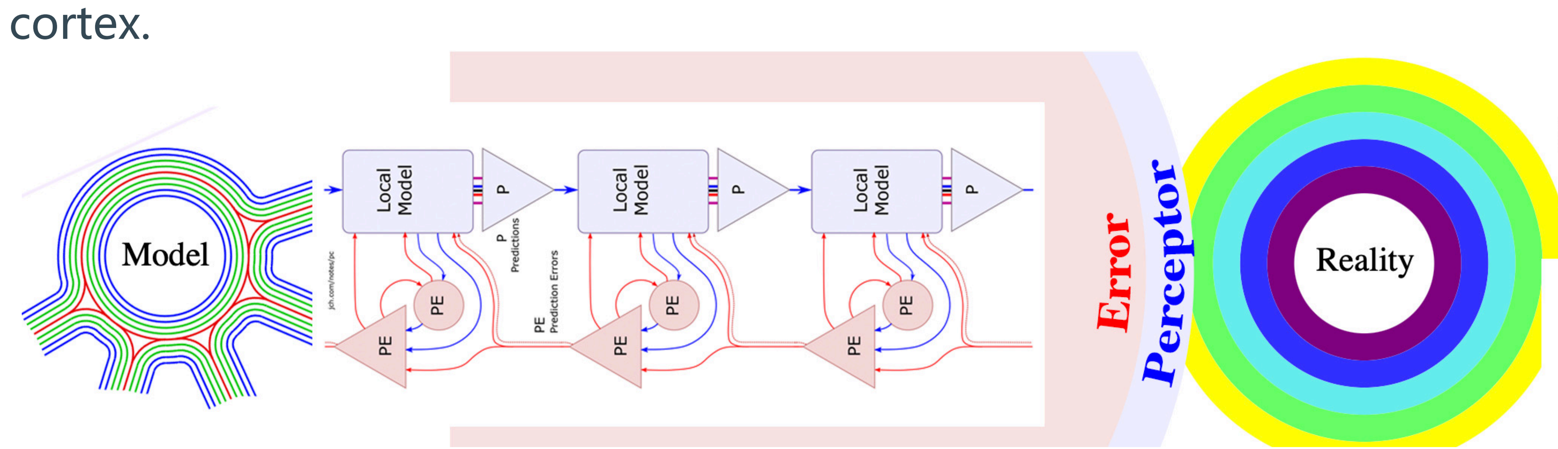
visual-language models with textual prompts

pure visual models with one normal image prompt

However, visual representations are inherently independent of language. In this study, we want to explore how far we can go with a pure visual model although there is room using visual-language models and worthwhile further to pursue.

2. Motivation

Predictive coding theory [1] postulates that the brain constantly generates and updates a "mental model". The mental model compares its expectations (or predictions) with the actual inputs from the visual cortex.



Some existing methods (i.e., PatchCore) perceive anomalies and they are indeed similar to the brains. However, they usually require a certain number of normal images and thus are limited in universal (i.e., open-world) scenarios.

We can build similar concepts in AS. First, given one normal image prompt for each class, we take it as the expected output. Then, the actual input could be any query images from the same class of the normal prompt. Last but not least, how to construct a "mental model" to compare between one normal image and these query images? Despite these challenges, we can imagine that the "mental model" should satisfy several basic principles.

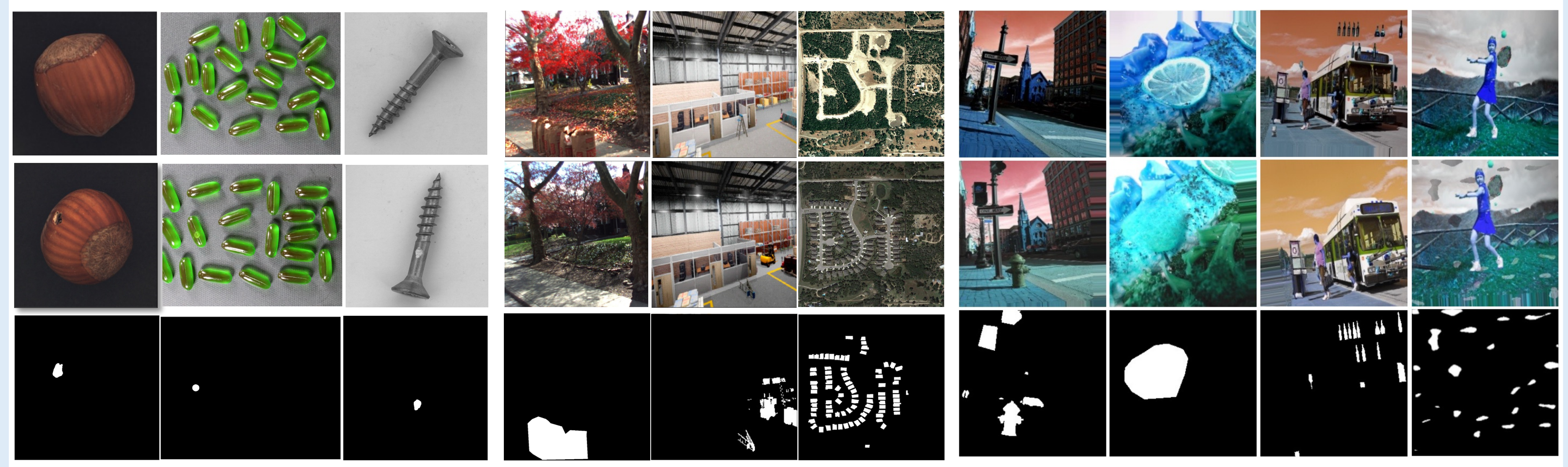
- ✓ First, it should have a strong generalization ability to perceive anomalies facing unseen objects or textures;
- ✓ Second, it can perform pixel-level anomaly segmentation only given one normal image prompt;
- ✓ Third, its training does not depend on target domain distribution or any guidance from language.

[1] https://jch.com/jch/notes/pc/

3. Our Method

3.1 Rethinking Anomaly Segmentation

We rethink anomaly segmentation and find it can be unified into change segmentation.

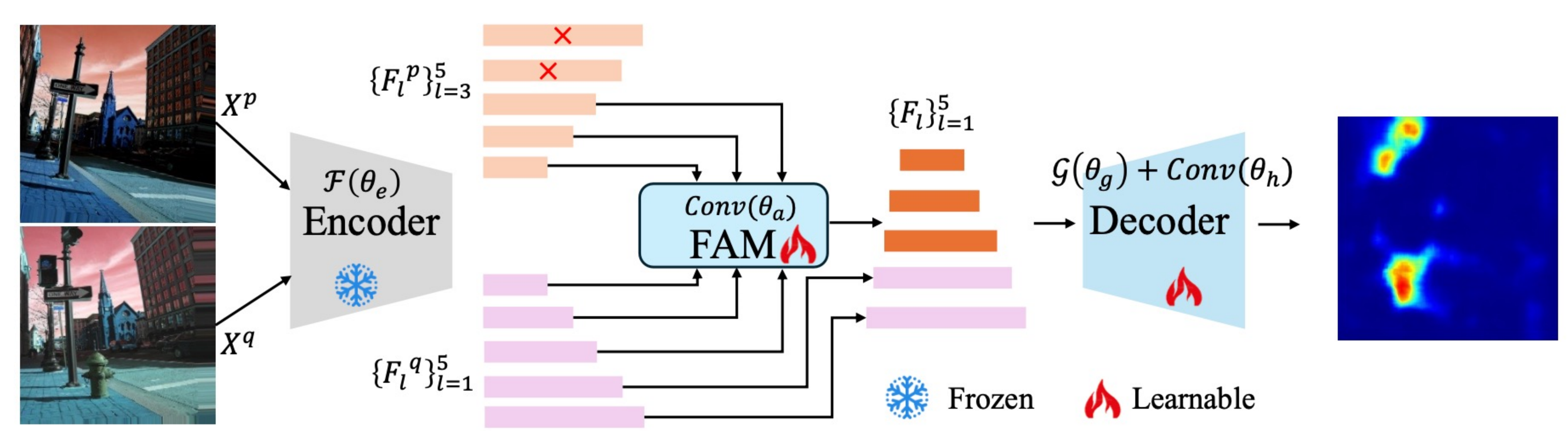


Anomaly Segmentation → Change Segmentation → Synthesizing Change Segmentation

It is not suitable using existing change segmentation datasets, such as street scenes, industrial environments, and remote sensing, for universal change segmentation due to small scale, insufficient diversity and various noises.

However, the novel paradigm shift enables us to leverage large-scale synthetic image pairs with **object-level** and **local region** changes, thereby overcoming the long-standing challenge of lacking large-scale anomaly segmentation datasets.

3.2 One-Prompt Meta-Learning for Universal Anomaly Segmentation



The proposed MetaUAS consists of an encoder, a feature alignment module (FAM), and a decoder. It is trained on a synthesized dataset in a one-prompt meta-learning manner for change segmentation tasks. Once trained, it can segment any anomalies providing only one normal image prompt.

Encoder: MetaUAS is compatible with any hierarchical architecture.

FAM aligns query and prompt features for better change segmentation.

$$F_l^p(i,j) \leftarrow F_l^p\left(\underset{k,l}{\operatorname{argmin}} \langle F_l^q(i,j), F_l^p(k,l)\rangle\right)$$

Hard Alignment

$$W_{ijkl} = \operatorname{softmax}\left(F_l^q(i,j)(F_l^p(k,l))^T\right),$$
$$F_l^p(i,j) \leftarrow \sum_k \sum_l W_{ijkl} F_l^p(k,l).$$

Soft Alignment

Decoder predicts each pixel to determine whether it is changed. Specifically, we utilize Unet as our decoder because it is better suited for tasks requiring high precision and the preservation of fine-grained details.

4. Experiments

4.1 Comparisons with the state-of-the-arts

Table 1: Quantitative comparisons on **MVTec**, **VisA** and **Goods**. Red indicates the best performance, while blue denotes the second-best result. Gray indicates the model is trained by full-shot normal images.

Datasets	Methods	Venue	Shot	Auxiliary	Anomaly Classification I-ROC	I-PR	$I\text{-}F1_{max}$	Anomaly Segmentation P-ROC	P-PR	$P\text{-}F1_{max}$	P-PRO
MVTec	CLIP [42]	ICML 21	0	✗	74.4	89.3	88.7	62.0	6.5	11.2	21.4
	PatchCore [47]	CVPR 22	1	✗	79.0±0.8	89.6±1.1	88.9±0.3	93.1±0.2	37.1±0.9	42.2±0.8	82.7±0.5
	WinCLIP [26]	CVPR 23	0	✗	90.4	95.6	92.7	82.3	18.2	24.8	61.9
	WinCLIP+ [26]	CVPR 23	1	✗	92.8±1.2	96.4±0.7	93.8±0.5	93.5±0.2	38.4±1.2	42.5±1.0	83.9±0.4
	AnomalyCLIP [76]	ICLR 24	0	✓	91.5	96.3	92.7	91.1	34.5	39.1	81.4
	UniAD [70]	NeurIPS 22	full	✗	96.7	98.9	96.7	96.8	44.7	50.4	90.0
	MetaUAS		1	✗	90.7±0.7	95.7±0.6	92.5±0.3	94.6±0.2	59.3±1.4	57.5±1.1	82.6±0.6
	MetaUAS⋆		1	✗	94.2	97.6	93.9	95.3	63.7	61.6	83.1
	MetaUAS⋆+		1	✗	95.3	97.9	94.6	97.6	67.0	62.9	92.5
VisA	CLIP [42]	ICML 21	0	✗	59.1	67.4	74.5	56.5	1.8	3.6	22.4
	PatchCore [47]	CVPR 22	1	✗	64.2±1.0	66.0±0.7	75.5±0.5	95.5±0.3	16.5±1.7	26.0±1.5	84.6±0.5
	WinCLIP [26]	CVPR 23	0	✗	75.5	78.7	78.2	73.2	5.4	9.0	51.0
	WinCLIP+ [26]	CVPR 23	1	✗	80.5±2.6	82.1±2.7	81.3±1.0	94.4±0.1	15.9±0.2	23.2±0.4	79.3±0.3
	AnomalyCLIP [76]	ICLR 24	0	✓	81.9	85.4	80.7	95.5	21.3	28.3	86.8
	UniAD [70]	NeurIPS 22	full	✗	90.8	93.2	87.8	98.5	34.3	39.1	84.8
	MetaUAS		1	✗	81.2±1.7	84.5±1.4	80.2±0.7	92.2±0.7	42.7±0.8	44.7±0.6	60.4±1.5
	MetaUAS⋆		1	✗	83.4	85.7	81.3	92.0	43.9	45.6	57.3
	MetaUAS⋆+		1	✗	85.1	87.2	82.3	98.0	48.1	48.6	85.5
Gooods	CLIP [42]	ICML 21	0	✗	51.8	57.3	71.3	55.3	4.3	2.0	16.4
	PatchCore [47]	CVPR 22	1	✗	48.3±1.0	54.2±0.5	71.3±0.1	84.3±0.5	4.5±0.2	9.3±0.3	55.6±1.0
	WinCLIP [26]	CVPR 23	0	✗	52.2	58.2	71.4	73.0	5.0	10.2	44.5
	WinCLIP+ [26]	CVPR 23	1	✗	53.5±0.2	58.6±0.2	71.5±0.1	85.5±0.6	5.7±0.4	11.3±0.5	56.6±1.2
	AnomalyCLIP [76]	ICLR 24	0	✓	57.2	63.3	71.4	83.5	16.9	24.0	63.3
	UniAD [70]	NeurIPS 22	full	✗	67.5	72.1	74.6	90.4	15.0	20.6	66.1
	MetaUAS		1	✗	54.5±1.0	58.5±0.4	71.5±0.1	88.5±0.6	8.6±0.7	14.0±0.7	59.0±1.3
	MetaUAS⋆		1	✗	90.1	91.7	85.7	97.4	53.7	55.5	70.8
	MetaUAS⋆+		1	✗	89.9	89.9	86.2	97.9	49.0	55.8	88.0

Table 2: The complexity and efficiency comparisons.

Methods	Backbone	#All Params(#Learnable)	Input Size	Times (ms)
CLIP [42]	ViT-B-16+240	208.4 (0.0)	240×240	13.7
PatchCore [47]	E-b4	17.5 (0.0)	256×256	36.4
			512×512	145.1
WinCLIP [26]	ViT-B-16+240	208.4 (0.0)	240×240	201.3
WinCLIP+ [26]				339.5
AnomalyCLIP [76]	ViT-L/14@336px	433.5 (5.6)	518×518	154.9
UniAD [70]	Eb4	27.1 (7.7)	224×224	5.0
MetaUAS	Eb4	22.1 (4.6)	256×256	3.1
MetaUAS⋆				
MetaUAS⋆+	Eb4+ViT-B-16+240	139.3 (4.6)		204.8
MetaUAS	Eb4	22.1 (4.6)	512×512	12.0
MetaUAS⋆				
MetaUAS⋆+	Eb4+ViT-B-16+240	139.3 (4.6)		213.0

- ✓ Strong generalization;
- ✓ High efficiency;
- ✓ Fewer parameters;
- ✓ Training-free on target domain;
- ✓ Only one normal image;
- ✓ Pure visual foundation model;
- ✓ Without any language prompt;

4.2 Ablation study

Table 3: **Ablation studies on MVTec**. Default settings are marked in blue.

(a) Effect of feature alignment module.

No.	Align	Fusion	I-ROC	I-PR	P-ROC	P-PR	P-PRO
1	No	Concat	82.8	92.5	88.4	44.9	67.5
2	Hard	Concat	87.1	94.7	90.7	48.2	77.0
3	Soft	Concat	**91.3**	**96.2**	**94.6**	**59.6**	**82.6**
4	Soft	Add	71.8	86.9	73.2	24.0	45.2
5	Soft	AbsDiff	84.1	92.4	88.4	45.9	68.4

(b) Learn or freeze encoder?

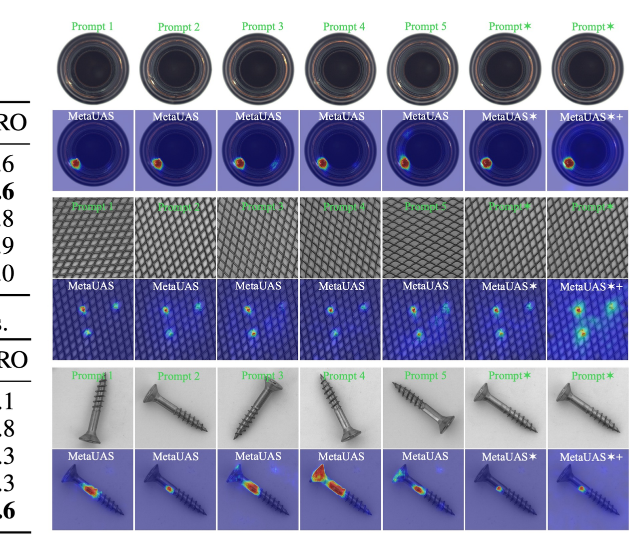
No.	Backbone	Learn?	I-ROC	I-PR	P-ROC	P-PR	P-PRO
1	E-b4	Learn	86.5	93.6	93.1	50.3	74.6
2	E-b4	Freeze	**91.3**	**96.2**	94.6	**59.6**	**82.6**
3	E-b6	Freeze	90.1	95.5	95.1	56.9	80.8
4	EViT-b3	Freeze	89.5	95.7	**95.3**	58.5	80.9
5	M-v2	Freeze	76.2	87.8	87.6	33.7	61.0

(c) Effects of change types and decoder module.

No.	ChangeType	Decoder	I-ROC	I-PR	P-ROC	P-PR	P-PRO
1	Only Loc.	UNet	83.1	92.8	87.7	44.3	76.1
2	Only Obj.	UNet	90.5	96.0	94.5	58.3	75.4
3	Obj.+Loc.	UNet	**91.3**	**96.2**	**94.6**	**59.6**	**82.6**
4	Obj.+Loc.	FPN-Cat	86.9	86.9	91.6	49.9	76.7
5	Obj.+Loc.	FPN-Add	88.4	94.7	94.1	51.4	73.1

(d) Effects of the number of training samples.

No.	#Samples	I-ROC	I-PR	P-ROC	P-PR	P-PRO
1	10%	82.0	91.9	85.4	36.5	62.1
2	30%	87.4	93.6	89.1	50.6	73.8
3	50%	91.0	96.2	92.9	57.1	74.3
4	70%	91.1	**96.4**	94.5	57.0	78.3
5	95%	**91.3**	96.2	**94.6**	**59.6**	**82.6**



Learn or freeze Encoder: learning the encoder will degenerate generalization.

FAM: soft alignment is better than no alignment and hard alignment.

Decoder: U-Net is better than FPN in pixel-level change segmentation.

ChangeTypes: The diversity of synthetic data is critical for good generalization.

Training Scale: MetaUAS still works when the number of training images is small scale (e.g., 50%), and the performance can further improve when increasing the number of training samples.

4.3 Qualitative Evaluation