FineCLIP: Self-distilled Region-based CLIP for Better Fine-grained Understanding

Dong Jing*, Xiaolong He*, Yutian Luo, Nanyi Fei,
Guoxing Yang, Wei Wei, Huiwen Zhao, Zhiwu Lu†

Outlines

1. Background
2. FineCLIP
3. In-domain Validation
4. Out-domain Evaluation with Scaled Trainset
5. Downstream Application
6. Conclusion

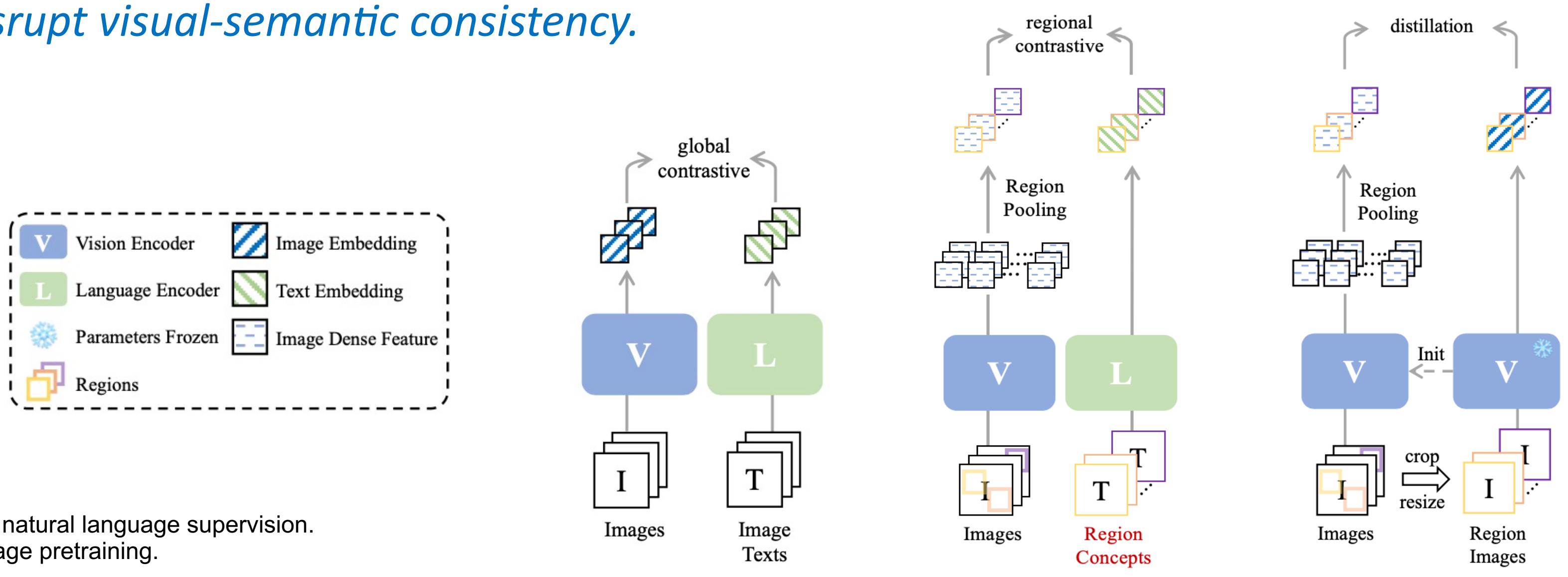
Background

1. CLIP[1] and its drawback

- CLIP aligns global image and text embeddings.
- Due to the weak supervision for visual dense features, CLIP performs poorly on dense prediction tasks

2. Two existing strategies for fine-grained alignment enhancement

- Matching image regions with template labels using large quantities of grounding annotations. [2,3]
  - Weakness: the pre-defined template labels lack sufficient semantic diversity.
- Global-to-region distillation with a frozen teacher. [4]
  - Weakness: the frozen teacher model restricts the performance ceiling of the student model.
- *Both disrupt visual-semantic consistency.*



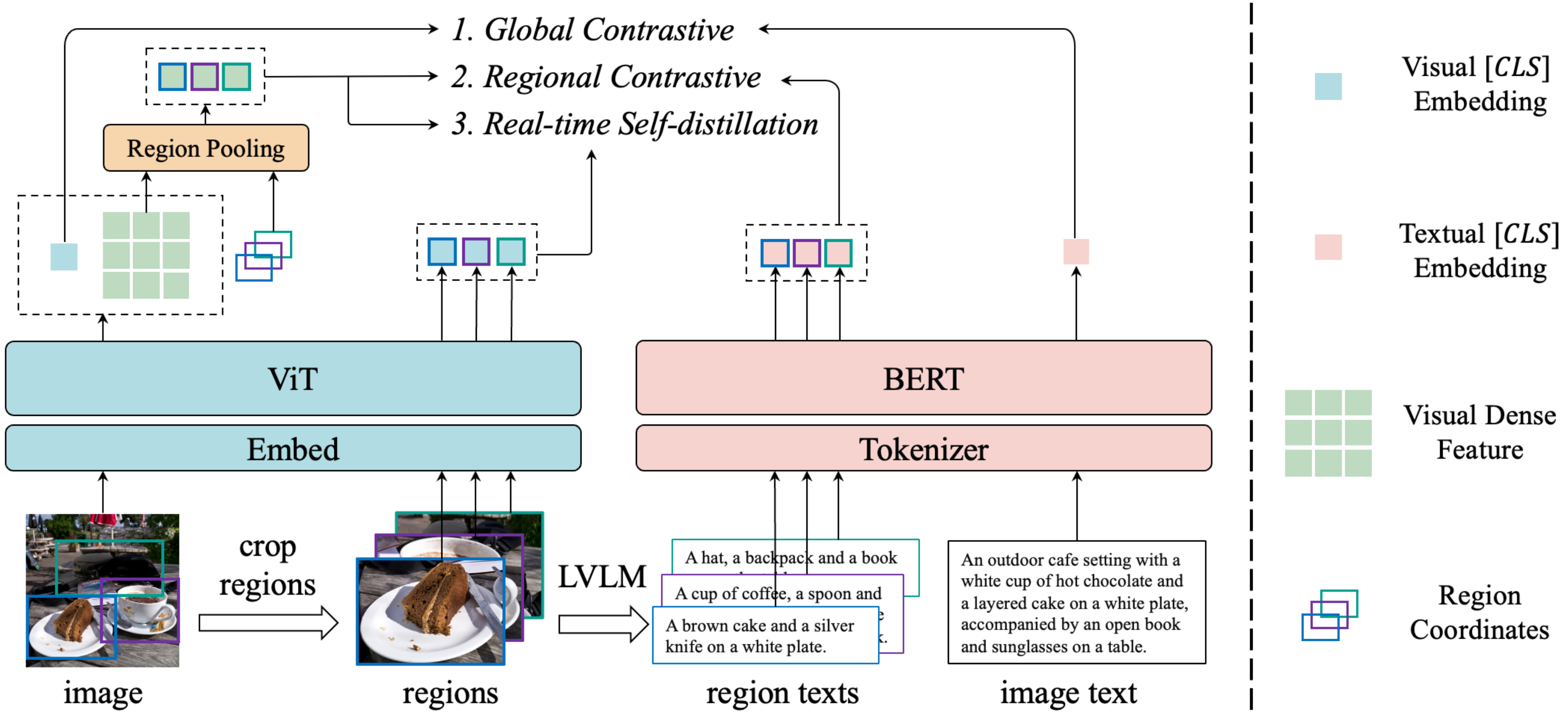
(a) CLIP (b) RegionCLIP (c) CLIPSelf

Reference:
[1] Learning transferable visual models from natural language supervision.
[2] RegionCLIP: Region-based language-image pretraining.
[3] Grounded language-image pre-training.
[4] CLIPSelf: Vision transformer distills itself for open-vocabulary dense prediction.

FineCLIP

FineCLIP incorporates THREE training components:

1. *Global Contrastive* – preserve global visual-semantic consistency & learn coarse knowledge from image-text pairs
2. *Regional Contrastive* – construct region-text alignment & learn fine-grained knowledge from region-text pairs
3. *Real-time Self-distillation* – interact the knowledge between region embeddings and pooled region features independently



Advantages of FineCLIP

- Diverse Semantics.
- No need for teacher, the model teach itself.
- Realize both global and regional visual-semantic alignments.

In-domain Validation

NEURAL INFORMATION PROCESSING SYSTEMS

高瓴人工智能学院 Gaoling School of Artificial Intelligence

1. Ablation Study

In-domain Setting:

- Train on COCO Train2017 split; Validate on COCO val2017 split[1]
- Model size: ViT-B/16; Input resolution: 224x224
- Region proposals are provided by COCO dataset
- Region captions are generated by BLIP-2 [2]

Ablation Results:

Table 1: Ablation study on the objective components.

#	L_{GC}	L_{SD}	L_{RC}	Box Classification Top1	Box Classification Top5	Retrieval I2T	Retrieval T2I
1		√		0.0	0.0	0.0	0.1
2	√			42.3	66.6	62.4	48.8
3	√	√		43.7	72.9	60.0	47.1
4			√	45.5	72.0	39.5	30.4
5	√		√	47.8	74.1	**62.5**	**48.9**
6	√	√	√	**48.4**	**75.6**	62.2	47.6

2. Comparisons with Competing Methods

Table 4: Performance comparisons of FineCLIP and competing methods on COCO.

#	Methods	Box Classification Top1	Box Classification Top5	Retrieval I2T	Retrieval T2I	Time Overhead (per epoch)	GPU Memory Usage (per card)
1	Pre-trained CLIP [40]	31.1	53.7	59.3	42.4	-	-
2	CLIP [40]	42.3	66.6	62.4	48.8	6 min	8G
3	RegionCLIP [70]	40.0	65.3	25.1	31.2	9 min	5G
4	CLIPSelf [55]	43.7	72.3	33.3	21.2	10 min	6G
5	FineCLIP(Ours)	48.4	75.6	62.2	47.6	11 min	36G

Reference:
[1] Microsoft coco: Common objects in context.
[2] Blip-2: Bootstrapping language- image pre-training with frozen image encoders and large language models.

Out-domain Evaluation on Scaled Trainset

高瓴人工智能学院
Gaoling School of Artificial Intelligence

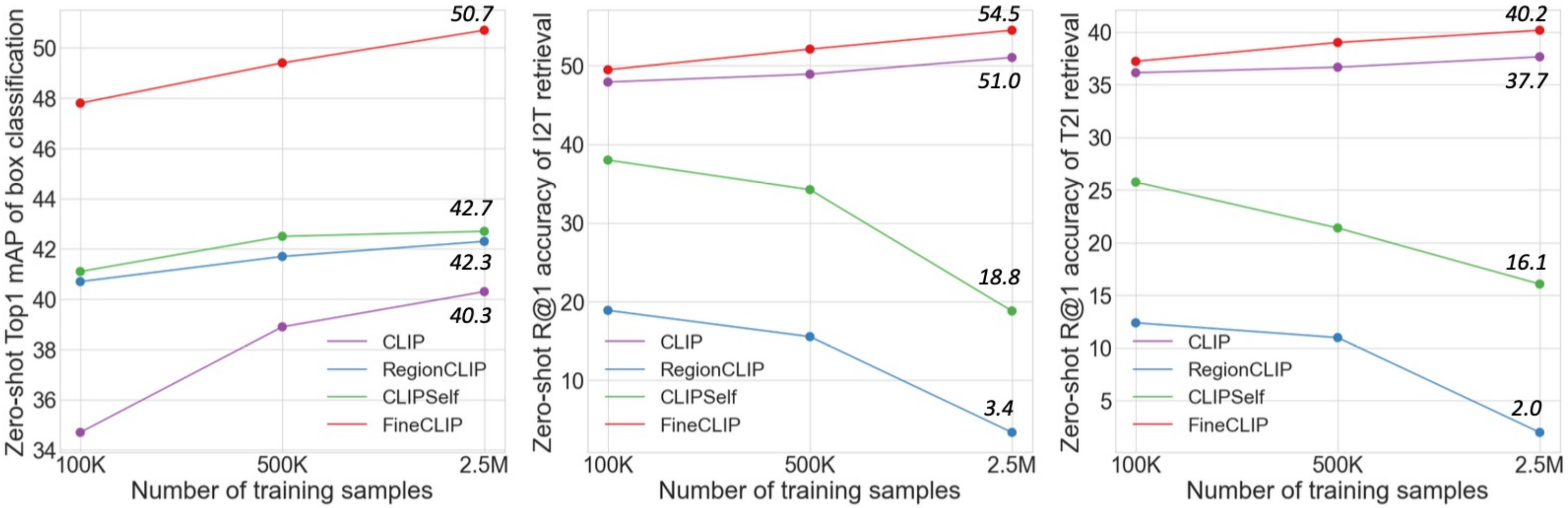
NEURAL INFORMATION PROCESSING SYSTEMS

1. Data Preparation based on CC3M [1]

- **Image Filtering**: we retain 2.5 million high-resolution images, referred to “CC2.5M”
- **Region Proposal**: we utilize YOLOv9 [2] to detect objects, which yields 10.4 million high-quality regions
- **Region Annotation**: we employ BLIP-2 [3] to annotate region proposals.

2. Out-domain Comparisons (Train on CC2.5M, Test on COCO)

FineCLIP presents
promising scalability



(a) Top1 mean accuracy of models on COCO box classification task.

(b) Accuracy of models on COCO R@1 image-to-text retrieval task.

(c) Accuracy of models on COCO R@1 text-to-image retrieval task.

Figure 2: Zero-shot comparisons of models pre-trained on datasets in three different scales.

Reference:
[1] Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning.
[2] Yolov9: Learning what you want to learn using programmable gradient information.
[3] Blip-2: Bootstrapping language- image pre-training with frozen image encoders and large language models.

Out-domain Evaluation on Scaled Trainset

高瓴人工智能学院
Gaoling School of Artificial Intelligence

NEURAL INFORMATION PROCESSING SYSTEMS

3. Visualization of attention maps of FineCLIP by GAE[1]

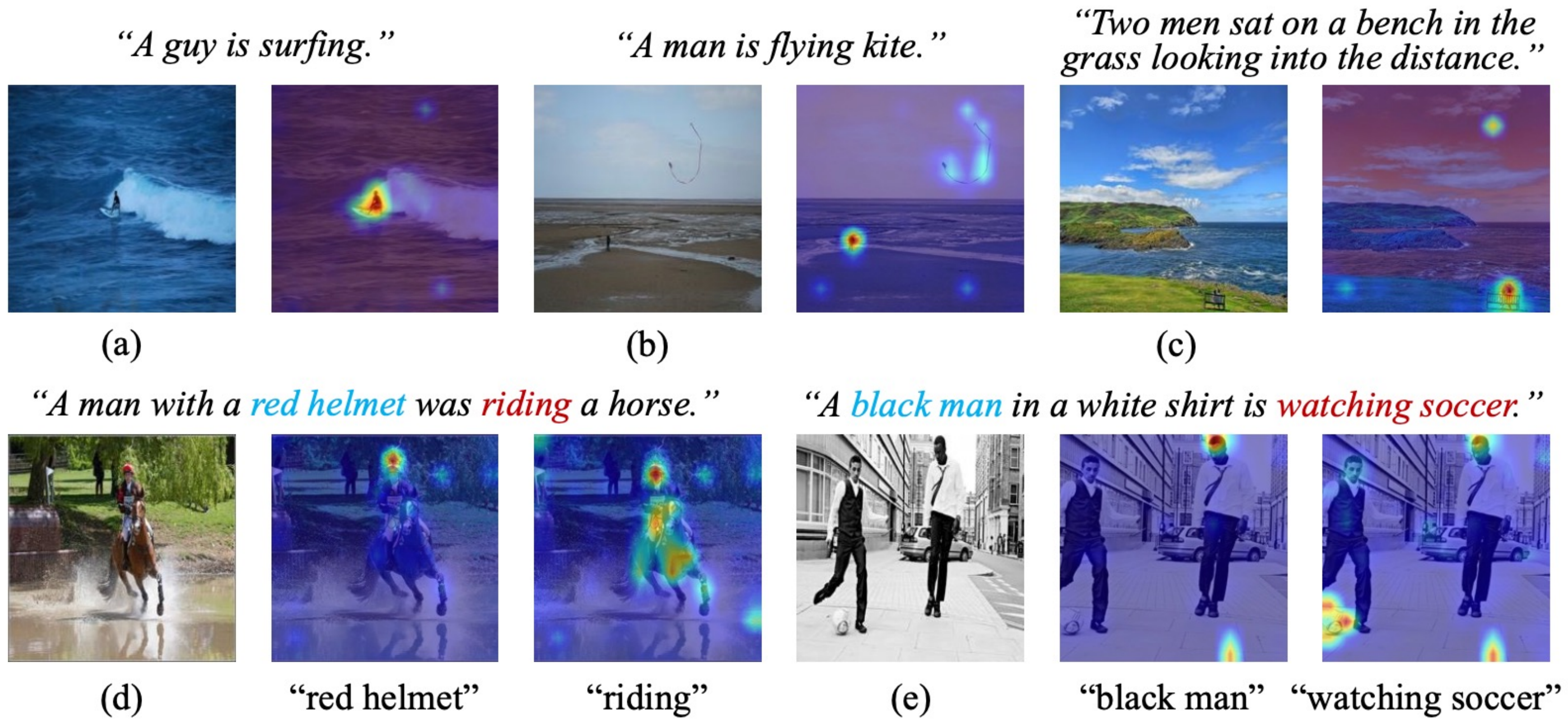


Figure 3: Visualizations of attention maps of our FineCLIP using GAE [5] on images responding to complete sentences or individual words. (a)-(c) Image attention maps w.r.t. different sentences. (d)(e) Image attention maps w.r.t. different words.

Reference:
[1] Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers.

Downstream Application

(Please refer to the original paper for detailed Settings)

1. Application to Fine-grained Localization

- **Open-vocabulary Object Detection**

(a) OV-COCO benchmark

Method	Backbone	AP_{50}^{novel}	AP_{50}^{base}	AP_{50}
OV-RCNN [63]	RN50	17.5	41.0	34.9
RegionCLIP [70]	RN50	26.8	54.8	47.5
PB-OVD [11]	RN50	30.8	46.1	42.1
Detic [73]	RN50	27.8	51.1	45.0
VLDet [29]	RN50	32.0	50.6	45.8
F-VLM [23]	RN50	28.0	-	39.6
BARON-Cap [54]	RN50	33.1	54.8	49.1
CORA [56]	RN50	35.1	35.5	35.4
RO-ViT [20]	ViT-B/16	30.2	-	41.5
RO-ViT [20]	ViT-L/16	33.0	-	47.7
CFM-ViT [19]	ViT-L/16	34.1	-	46.0
F-ViT	ViT-B/16	17.5	41.0	34.9
F-ViT+CLIPSelf†	ViT-B/16	25.4	40.9	36.8
F-ViT+FineCLIP†	ViT-B/16	29.8↑12.3	45.9↑4.9	41.7↑6.8
F-ViT	ViT-L/14	24.7	53.6	46.0
F-ViT+CLIPSelf†	ViT-L/14	38.4	54.4	50.2
F-ViT+FineCLIP†	ViT-L/14	**40.0**↑15.3	**57.2**↑3.6	**52.7**↑6.7

- **Open-vocabulary Semantic Segmentation**

Table 6: Results on open-vocabulary semantic segmentation. † means the CLIP ViT backbone is initialized with the checkpoint of the corresponding method trained on CC2.5M.

Method	Backbone	ADE-150 mIoU	ADE-150 mAcc	ADE-847 mIoU	ADE-847 mAcc	PC-59 mIoU	PC-59 mAcc
OVSeg [28]	ViT-B/16	24.8	-	7.1	-	53.3	-
SAN [57]	ViT-B/16	27.5	45.6	10.1	21.1	53.8	73.0
SAN [57]	ViT-L/14	32.1	50.7	12.4	**25.2**	57.7	77.6
CatSeg [7]	ViT-B/16	27.2	41.2	8.4	16.6	57.5	74.0
CatSeg [7]	ViT-L/14	31.5	46.2	10.8	20.5	**62.0**	**78.3**
CatSeg+CLIPSelf† [55]	ViT-B/16	29.7	45.1	10.1	17.2	55.3	73.4
CatSeg+CLIPSelf† [55]	ViT-L/14	34.9	52.9	13.6	23.0	59.1	77.1
CatSeg+FineCLIP†	ViT-B/16	32.4↑5.2	50.5↑9.3	12.2↑4.2	22.2↑5.6	56.0↓1.5	74.4↑0.4
CatSeg+FineCLIP†	ViT-L/14	**36.1**↑4.6	**53.5**↑7.3	**14.1**↑3.3	23.8↑3.3	59.9↓2.1	**78.3**↑0

Downstream Application

(Please refer to the original paper for detailed Settings)

2. Application to Image-level Task

- **Zero-shot image-text retrieval**

Table 7: Comparative results on zero-shot image-text retrieval on the Flickr30k and MSCOCO datasets. R@i denotes Recall at i. All approaches adopt ViT-B/16 architecture with input image size of 224 $\times$ 224. † indicates that the method is initialized with pre-trained CLIP and further trained on CC2.5M. The methods with gray background are pre-trained on large-scale dataset.

	Flickr30k						MSCOCO					
	image-to-text			text-to-image			image-to-text			text-to-image		
Methods	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP[40]	84.0	96.1	98.2	71.6	90.3	94.1	56.2	80.6	88.2	42.4	68.6	78.3
SPARC[3]	84.4	97.6	98.7	72.0	91.2	94.9	57.6	81.2	88.5	43.0	68.6	78.5
PACL[35]	69.6	89.7	94.2	54.9	80.7	87.3	41.8	67.8	77.6	29.1	54.3	65.5
GLoRIA[16]	78.0	95.5	98.0	68.4	88.9	93.2	49.7	75.4	84.6	38.9	65.1	75.2
MGCA[52]	82.2	96.1	98.1	67.7	88.5	93.2	57.6	80.5	87.8	39.8	65.7	75.3
FILIP[58]	69.0	89.8	94.0	55.8	81.5	87.9	40.2	66.0	76.3	29.5	55.3	66.3
CLIP† [40]	81.6	96.2	98.0	64.9	88.3	93.6	51.1	76.4	84.9	37.6	63.9	74.3
RegionCLIP†[70]	3.9	12.2	18.4	7.9	22.7	71.3	2.0	7.1	11.5	3.4	11.8	19.0
CLIPSelf†[55]	33.8	61.7	73.0	35.0	61.3	32.7	18.8	38.9	50.4	16.1	34.5	45.1
FineCLIP†	**82.5**	**96.4**	**98.6**	**67.9**	**89.1**	**94.1**	**54.5**	**78.6**	**85.8**	**40.2**	**66.5**	**76.1**

Conclusion

- **We present FineCLIP**, which combines multi-grained contrastive learning paradigm and the real-time self-distillation scheme to achieve better fine-grained understanding.

- **We develop an automated region-text data generation pipeline** utilizing advanced LVLMs, and demonstrate its effectiveness in providing valuable fine-grained semantics.

- Extensive experiments on dense prediction and image-level benchmarks show that our **FineCLIP outperforms previous arts in most scenes and exhibits promising scalability.**