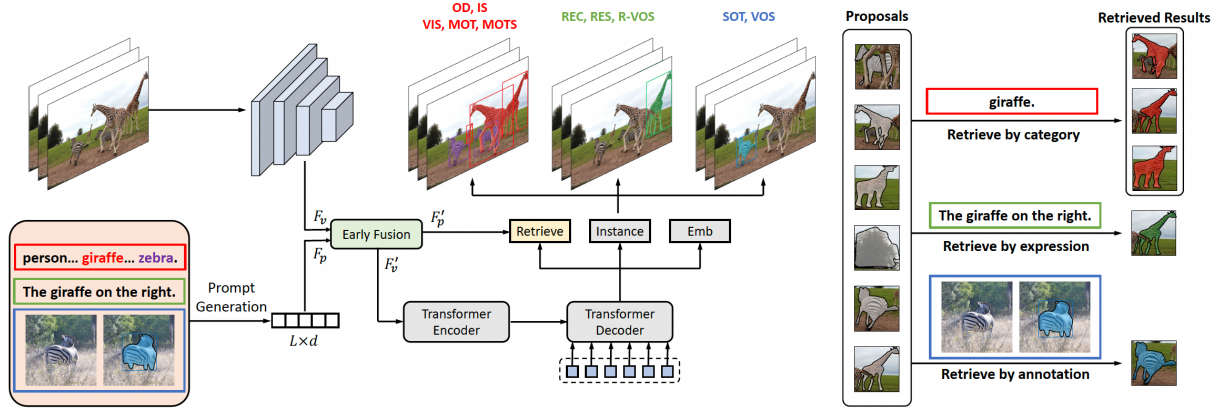


Universal Instance Perception as Object Discovery and Retrieval



This is the official implementation of the paper Universal Instance Perception as Object Discovery and Retrieval.

	Ranked #4	Referring Expression Comprehension on RefCoco+ (using additional training data)
	Ranked #4	Referring Expression Comprehension on RefCoco+ (using additional training data)
	Ranked #4	Referring Expression Comprehension on RefCoco+ (using additional training data)
	Ranked #4	Referring Expression Comprehension on RefCoco+ (using additional training data)
	Ranked #4	Referring Expression Comprehension on RefCoco+ (using additional training data)
	Ranked #4	Referring Expression Comprehension on RefCoco+ (using additional training data)
	Ranked #4	Referring Expression Comprehension on RefCoco+ (using additional training data)
	Ranked #4	Referring Expression Comprehension on RefCoco+ (using additional training data)
	Ranked #4	Referring Expression Comprehension on RefCoco+ (using additional training data)
	Ranked #4	Referring Expression Comprehension on RefCoco+ (using additional training data)
	Ranked #4	Referring Expression Comprehension on RefCoco+ (using additional training data)
	Ranked #4	Referring Expression Comprehension on RefCoco+ (using additional training data)
	Ranked #4	Referring Expression Comprehension on RefCoco+ (using additional training data)
	Ranked #4	Referring Expression Comprehension on RefCoco+ (using additional training data)
	Ranked #4	Referring Expression Comprehension on RefCoco+ (using additional training data)

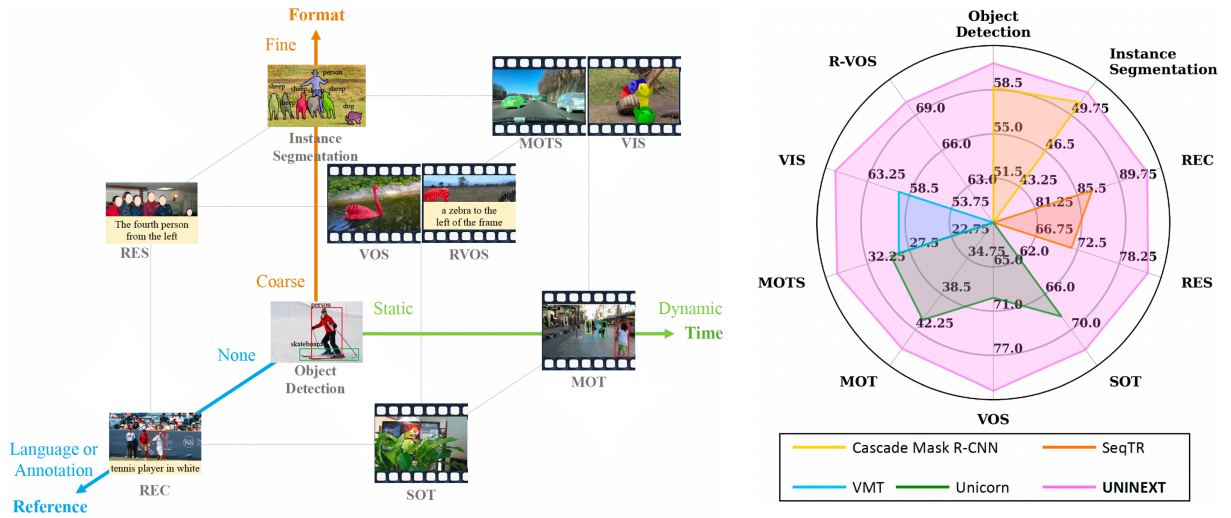
News

- :trophy: We are the runner-up in Segmentation in the Wild challenge.
- :trophy: We are the winner of BDD100K MOT Challenge and the runner-up of BDD MOTS Challenge on CVPR2023 workshop.

Highlight

- UNINEXT is accepted by **CVPR2023**.
- UNINEXT reformulates diverse instance perception tasks into **a unified object discovery and retrieval paradigm** and can flexibly perceive different types of objects by simply changing the input prompts.
- UNINEXT achieves **superior performance on 20 challenging benchmarks using a single model with the same model parameters**.

Introduction



Object-centric understanding is one of the most essential and challenging problems in computer vision. In this work, we mainly discuss 10 sub-tasks, distributed on the vertices of the cube shown in the above figure. Since all these tasks aim to perceive instances of certain properties, UNINEXT reorganizes them into three types according to the different input prompts: - Category Names - Object Detection - Instance Segmentation - Multiple Object Tracking (MOT) - Multi-Object Tracking and Segmentation (MOTS) - Video Instance Segmentation (VIS) - Language Expressions - Referring Expression Comprehension (REC) - Referring Expression Segmentation (RES) - Referring Video Object Segmentation (R-VOS) - Target Annotations - Single Object Tracking (SOT) - Video Object Segmentation (VOS)

Then we propose a unified prompt-guided object discovery and retrieval formulation to solve all the above tasks. Extensive experiments demonstrate that UNINEXT achieves superior performance on 20 challenging benchmarks.

Demo

<https://user-images.githubusercontent.com/40926230/224527028-f31e8de0-b8aa-4cfb-a83b-63a70ff5bd52.mp4>

UNINEXT can flexibly perceive various types of objects by simply changing the input prompts, such as category names, language expressions, and target annotations. We also provide a simple demo script, which supports 4 image-level tasks (object detection, instance segmentation, REC, RES).

Results

Retrieval by Category Names

Object Detection (COCO 2017 val)

Model	Backbone	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
Faster R-CNN [87]	ResNet-50	42.0	62.1	45.5	26.6	45.4	53.4
DETR [9]		43.3	63.1	45.9	22.5	47.3	61.1
Sparse R-CNN [93]		45.0	63.4	48.2	26.9	47.2	59.5
Cascade Mask-RCNN [8]		46.3	64.3	50.5	-	-	-
Deformable-DETR [135]		46.9	65.6	51.0	29.6	50.1	61.6
DN-Deformable-DETR [55]		48.6	67.4	52.7	31.0	52.0	63.7
UNINEXT		51.3	68.4	56.2	32.6	55.7	66.5
HTC++ [12]	Swin-L	58.0	-	-	-	-	-
DyHead [21]		60.3	-	-	-	-	-
Cascade Mask R-CNN [8]	ConvNeXt-L	54.8	73.8	59.8	-	-	-
UNINEXT		58.1	74.9	63.7	40.7	62.5	73.6
ViTDet-H [74]	ViT-H	58.7	-	-	-	-	-
UNINEXT		60.6	77.5	66.7	45.1	64.8	75.3

Instance Segmentation(COCO 2017 test-dev)

Model	Backbone	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
CondInst [94]	ResNet-50	38.6	60.2	41.4	20.6	41.0	51.1
Cascade Mask R-CNN [8]		38.6	60.0	41.7	21.7	40.8	49.6
SOLOv2 [103]		38.8	59.9	41.7	16.5	41.7	56.2
HTC [12]		39.7	61.4	43.1	22.6	42.2	50.6
QueryInst [31]		40.6	63.0	44.0	23.4	42.5	52.8
UNINEXT		44.9	67.0	48.9	26.3	48.5	59.0
QueryInst [31]	Swin-L	49.1	74.2	53.8	31.5	51.8	63.2
Mask2Former [16]*		50.1	-	-	29.9	53.9	72.1
Cascade Mask R-CNN [8]	ConvNeXt-L	47.6	71.3	51.7	-	-	-
UNINEXT		49.6	73.4	54.3	30.4	53.6	65.7
ViTDet-H [74]*	ViT-H	50.9	-	-	-	-	-
UNINEXT		51.8	76.2	56.7	33.3	55.9	67.5

Video Instance Segmentation

Method	Backbone	Online	VIS2019 val			OVIS val		
			AP	AP ₅₀	AP ₇₅	AP	AP ₅₀	AP ₇₅
VisTR [104]	ResNet-50	✗	36.2	59.8	36.9	-	-	-
MaskProp [4]		✗	40.0	-	42.9	-	-	-
IFC [45]		✗	42.8	65.8	46.8	13.1	27.8	11.6
SeqFormer [107]		✗	47.4	69.8	51.8	15.1	31.9	13.8
IDOL [109]		✓	49.5	74.0	52.9	30.2	51.3	30.0
VITA [41]		✗	49.8	72.6	54.5	19.6	41.2	17.4
UNINEXT		✓	53.0	75.2	59.1	34.0	55.5	35.6
SeqFormer [107]	Swin-L	✗	59.3	82.1	66.4	-	-	-
VMT [48]		✗	59.7	-	66.7	19.8	39.6	17.2
VITA [41]		✗	63.0	86.9	67.9	-	-	-
IDOL [109]		✓	64.3	87.5	71.0	42.6	65.7	45.2
UNINEXT	ConvNeXt-L	✓	64.3	87.2	71.7	41.1	65.8	42.0
UNINEXT	ViT-H	✓	66.9	87.5	75.1	49.0	72.5	52.2

MOT (BDD100K)

Method	Split	mMOTA↑	mIDF1↑	MOTA↑	IDF1↑	ID Sw.↓
Yu <i>et al.</i> [124]	val	25.9	44.5	56.9	66.8	8315
QDTrack [82]	val	36.6	50.8	63.5	71.5	6262
Unicorn [113]	val	41.2	54.0	66.6	71.3	10876
UNINEXT-L	val	41.8	54.9	64.6	68.7	9134
UNINEXT-H	val	44.2	56.7	67.1	69.9	10222

MOTS (BDD100K)

Method	Online	mMOTSA↑	mMOTSP↑	mIDF1↑	ID Sw.↓
SortIoU	✓	10.3	59.9	21.8	15951
MaskTrackRCNN [116]	✓	12.3	59.9	26.2	9116
STEm-Seg [1]	✗	12.2	58.2	25.4	8732
QDTrack-mots [82]	✓	22.5	59.6	40.8	1340
PCAN [49]	✓	27.4	66.7	45.1	876
VMT [48]	✗	28.7	67.3	45.7	825
Unicorn [113]	✓	29.6	67.7	44.2	1731
UNINEXT-L	✓	32.0	60.2	45.4	1634
UNINEXT-H	✓	35.7	68.1	48.5	1776

REC (Prec@0.5)										RES (on			
Method	RefCOCO			RefCOCO+			RefCOCOg		Method	RefCOCO			
	val	testA	testB	val	testA	testB	val-u	test-u		val	testA	testB	
UNITER _L [15]	81.41	87.04	74.17	75.90	81.45	66.70	74.86	75.77	CMSA [123]	58.32	60.61	55.09	
VILLA _L [34]	82.39	87.48	74.84	76.17	81.54	66.84	76.18	76.71	BRINet [43]	60.98	62.99	59.21	
MDETR [47]	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89	CMPC+ [65]	62.47	65.08	60.82	
RefTR [57]	85.65	88.73	81.16	77.55	82.26	68.99	79.25	80.01	MCN [72]	62.44	64.20	59.71	
SeqTR [134]	87.00	90.15	83.59	78.69	84.51	71.87	82.69	83.37	EFN [33]	62.76	65.69	59.67	
UNINEXT-R50	89.72	91.52	86.93	79.76	85.23	72.78	83.95	84.31	VLT [26]	65.65	68.29	62.73	
UNINEXT-L	91.43	93.73	88.93	83.09	87.90	76.15	86.91	87.48	SeqTR [134]	71.70	73.31	69.82	
UNINEXT-H	92.64	94.33	91.46	85.24	89.63	79.79	88.73	89.37	LAVT [120]	72.73	75.82	68.79	
									UNINEXT-R50	77.90	79.68	75.77	
									UNINEXT-L	80.32	82.61	77.76	
									UNINEXT-H	82.19	83.44	81.33	

Retrieval by Language Expressions

R-VOS									
Method	Backbone	Ref-Youtube-VOS			Ref-DAVIS17				
		$\mathcal{J} \& \mathcal{F}$	\mathcal{J}	\mathcal{F}	$\mathcal{J} \& \mathcal{F}$	\mathcal{J}	\mathcal{F}		
CMSA [123]	ResNet-50	36.4	34.8	38.1	40.2	36.9	43.5		
URVOS [90]		47.2	45.3	49.2	51.5	47.3	56.0		
YOFO [54]		48.6	47.5	49.7	54.4	50.1	58.7		
ReferFormer [108]		58.7	57.4	60.1	58.5	55.8	61.3		
UNINEXT		61.2	59.3	63.0	63.9	59.6	68.1		
PMINet + CFBI [27]	Ensemble	54.2	53.0	55.5	-	-	-		
CITD [58]		61.4	60.0	62.7	-	-	-		
MTTR [7]	Video-Swin-T	55.3	54.0	56.6	-	-	-		
ReferFormer [108]		64.9	62.8	67.0	61.1	58.1	64.1		
UNINEXT	ConvNext-L	66.2	64.0	68.4	66.7	62.3	71.1		
UNINEXT		70.1	67.6	72.7	72.5	68.2	76.8		

SOT											
Method	Backbone	LaSOT [30]			LaSOT _{ext} [29]			TrackingNet [79]			TNL-2K
		AUC	P _{Norm}	P	AUC	P _{Norm}	P	AUC	P _{Norm}	P	AUC
PrDiMP [22]	ResNet-50	59.8	68.8	60.8	-	-	-	75.8	81.6	70.4	47.0
LTMU [20]		57.2	-	57.2	41.4	49.9	47.3	-	-	-	48.5
TransT [14]		64.9	73.8	69.0	-	-	-	81.4	86.7	80.3	50.7
KeepTrack [75]		67.1	77.2	70.2	48.2	-	-	-	-	-	-
UNINEXT		69.2	77.1	75.5	51.2	58.1	58.1	83.2	86.9	83.3	56.0
SimTrack [10]	ViT-B	69.3	78.5	-	-	-	-	82.3	-	86.5	54.8
OSTrack [122]		71.1	81.1	77.6	50.5	61.3	57.6	83.9	88.5	83.2	55.9
Unicorn [113]	ConvNeXt-L	68.5	76.6	74.1	-	-	-	83.0	86.4	82.2	-
UNINEXT		72.4	80.7	78.9	54.4	61.8	61.4	85.1	88.2	84.7	58.1
UNINEXT	ViT-H	72.2	80.7	79.4	56.2	63.8	63.8	85.4	89.0	86.4	59.3

Retrieval by Target Annotations

VOS									
Method		YT-VOS 2018 val [111]					DAVIS 2017 val [83]		
		\mathcal{G}	\mathcal{J}_s	\mathcal{F}_s	\mathcal{J}_u	\mathcal{F}_u	$\mathcal{J} \& \mathcal{F}$	\mathcal{J}	\mathcal{F}
Memory	STM [81]	79.4	79.7	84.2	72.8	80.9	81.8	79.2	84.3
	CFBI [121]	81.4	81.1	85.8	75.3	83.4	81.9	79.1	84.6
	STCN [18]	83.0	81.9	86.5	77.9	85.7	85.4	82.2	88.6
	XMem [17]	86.1	85.1	89.8	80.3	89.2	87.7	84.0	91.4
Non-Memory	SiamMask [100]	52.8	60.2	58.2	45.1	47.7	56.4	54.3	58.5
	Unicorn [113]	-	-	-	-	-	69.2	65.2	73.2
	Siam R-CNN [98]	73.2	73.5	-	66.2	-	70.6	66.1	75.0
	TVOS [130]	67.8	67.1	69.4	63.0	71.6	72.3	69.9	74.7
	FRTM [89]	72.1	72.3	76.2	65.9	74.1	76.7	73.9	79.6
	UNINEXT-R50	77.0	76.8	81.0	70.8	79.4	74.5	71.3	77.6
	UNINEXT-L	78.1	79.1	83.5	71.0	78.9	77.2	73.2	81.2
	UNINEXT-H	78.6	79.9	84.9	70.6	79.2	81.8	77.7	85.8

Getting started

1. Installation: Please refer to INSTALL.md for more details.
2. Data preparation: Please refer to DATA.md for more details.
3. Training: Please refer to TRAIN.md for more details.
4. Testing: Please refer to TEST.md for more details.

-
5. Model zoo: Please refer to MODEL_ZOO.md for more details.

Citing UNINEXT

If you find UNINEXT useful in your research, please consider citing:

```
1 @inproceedings{UNINEXT,  
2   title={Universal Instance Perception as Object Discovery and  
3     Retrieval},  
4   author={Yan, Bin and Jiang, Yi and Wu, Jiannan and Wang, Dong and  
5     Yuan, Zehuan and Luo, Ping and Lu, Huchuan},  
6   booktitle={CVPR},  
   year={2023}
```

Acknowledgments

- Thanks Unicorn for providing experience of unifying four object tracking tasks (SOT, MOT, VOS, MOTS).
- Thanks VNext for providing experience of Video Instance Segmentation (VIS).
- Thanks ReferFormer for providing experience of REC, RES, and R-VOS.
- Thanks GLIP for the idea of unifying object detection and phrase grounding.
- Thanks Detic for the implementation of multi-dataset training.
- Thanks detrex for the implementation of denoising mechanism.