Universal Instance Perception as Object Discovery and Retrieval

Bin Yan[1]*, Yi Jiang[2,†], Jiannan Wu[3], Dong Wang[1],
Ping Luo[3], Zehuan Yuan[2], Huchuan Lu[1,4,†]
[1] School of Information and Communication Engineering, Dalian University of Technology, China
[2] ByteDance [3] The University of Hong Kong [4] Peng Cheng Laboratory

Abstract

All instance perception tasks aim at finding certain objects specified by some queries such as category names, language expressions, and target annotations, but this complete field has been split into multiple independent subtasks. In this work, we present a ***uni****versal* ***in****stance perception model of the* ***next*** *generation, termed* ***UNINEXT****. 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. This unified formulation brings the following benefits: (1) enormous data from different tasks and label vocabularies can be exploited for jointly training general instance-level representations, which is especially beneficial for tasks lacking in training data. (2) the unified model is parameter-efficient and can save redundant computation when handling multiple tasks simultaneously. UNINEXT shows superior performance on 20 challenging benchmarks from 10 instance-level tasks including classical image-level tasks (object detection and instance segmentation), vision-and-language tasks (referring expression comprehension and segmentation), and six video-level object tracking tasks. Code is available at https://github.com/MasterBin-IIAU/UNINEXT.*

1. Introduction

Object-centric understanding is one of the most essential and challenging problems in computer vision. Over the years, the diversity of this field increases substantially. In this work, we mainly discuss 10 sub-tasks, distributed on the vertices of the cube shown in Figure 1. As the most fundamental tasks, object detection [8, 9, 30, 59, 82, 84, 91] and instance segmentation [6, 37, 64, 90, 97] require finding all objects of specific categories by boxes and masks respectively. Extending inputs from static images to dynamic videos, Multiple Object Tracking (MOT) [3, 74, 126, 128], Multi-Object Tracking and Segmentation (MOTS) [47, 93, 108], and Video Instance Segmentation (VIS) [43, 100, 103, 112] require finding all object trajectories of specific categories in videos. Except for **category names**, some tasks provide other reference information. For example, Referring Expression Comprehension (REC) [115, 121, 130], Referring Expression Segmentation (RES) [116, 119, 121], and Referring Video Object Segmentation (R-VOS) [7, 86, 104] aim at finding objects matched with the given **language expressions** like "The fourth person from the left". Besides, Single Object Tracking (SOT) [5, 51, 106] and Video Object Segmentation (VOS) [18, 78, 107] take the **target annotations** (boxes or masks) given in the first frame as the reference, requiring to predict the trajectories of the tracked objects in the subsequent frames. Since all the above tasks aim to perceive instances of certain properties, we refer to them collectively as *instance perception*.

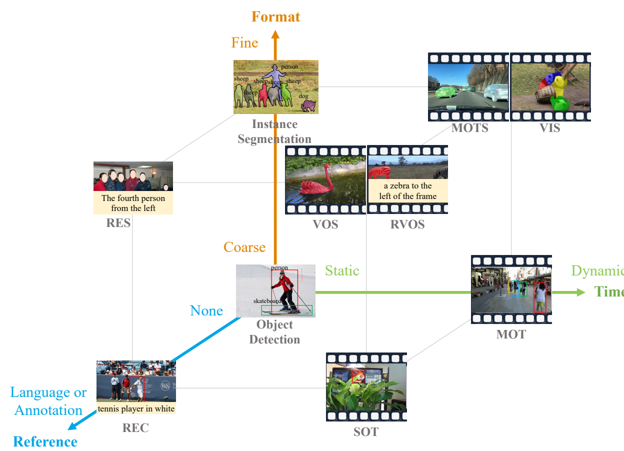

Figure 1. Task distribution on the Format-Time-Reference space. Better view on screen with zoom-in.

Although bringing convenience to specific applications, such diverse task definitions split the whole field into fragmented pieces. As the result, most current instance perception methods are developed for only a single or a part of sub-tasks and trained on data from specific domains. Such

*This work was performed while Bin Yan worked as an intern at ByteDance. Email: yan_bin@mail.dlut.edu.cn. † Corresponding authors: jiangyi.enjoy@bytedance.com, lhchuan@dlut.edu.cn.

fragmented design philosophy brings the following drawbacks: (1) Independent designs hinder models from learning and sharing generic knowledge between different tasks and domains, causing redundant parameters. (2) The possibility of mutual collaboration between different tasks is overlooked. For example, object detection data enables models to recognize common objects, which can naturally improve the performance of REC and RES. (3) Restricted by fixed-size classifiers, traditional object detectors are hard to jointly train on multiple datasets with different label vocabularies [36,61,88] and to dynamically change object categories to detect during inference [22, 61, 74, 81, 112, 120]. Since *essentially all instance perception tasks aim at finding certain objects according to some queries*, it leads to a natural question: could we design a unified model to solve all mainstream instance perception tasks once and for all?

To answer this question, we propose UNINEXT, a universal instance perception model of the next generation. We first reorganize 10 instance perception tasks into three types according to the different input prompts: (1) **category names as prompts** (Object Detection, Instance Segmentation, VIS, MOT, MOTS). (2) **language expressions as prompts** (REC, RES, R-VOS). (3) **reference annotations as prompts** (SOT, VOS). Then we propose a unified prompt-guided object discovery and retrieval formulation to solve all the above tasks. Specifically, **UNINEXT first discovers N object proposals under the guidance of the prompts, then retrieves the final instances from the proposals according to the instance-prompt matching scores**. Based on this new formulation, UNINEXT can flexibly perceive different instances by simply changing the input prompts. To deal with different prompt modalities, we adopt a prompt generation module, which consists of a reference text encoder and a reference visual encoder. Then an early fusion module is used to enhance the raw visual features of the current image and the prompt embeddings. This operation enables deep information exchange and provides highly discriminative representations for the later instance prediction step. Considering the flexible query-to-instance fashion, we choose a Transformer-based object detector [131] as the instance decoder. Specifically, the decoder first generates N instance proposals, then the prompt is used to retrieve matched objects from these proposals. This flexible retrieval mechanism overcomes the disadvantages of traditional fixed-size classifiers and enables joint training on data from different tasks and domains.

With the unified model architecture, UNINEXT can learn strong generic representations on massive data from various tasks and solve 10 instance-level perception tasks using a single model with the same model parameters. Extensive experiments demonstrate that UNINEXT achieves superior performance on 20 challenging benchmarks. The contributions of our work can be summarized as follows.

- We propose a unified prompt-guided formulation for universal instance perception, reuniting previously fragmented instance-level sub-tasks into a whole.
- Benefiting from the flexible object discovery and retrieval paradigm, UNINEXT can train on different tasks and domains, in no need of task-specific heads.
- UNINEXT achieves superior performance on 20 challenging benchmarks from 10 instance perception tasks using a single model with the same model parameters.

2. Related Work

Instance Perception. The goals and typical methods of 10 instance perception tasks are introduced as follows.

Retrieval by Category Names. Object detection and instance segmentation aim at finding all objects of specific classes on the images in the format of boxes or masks. Early object detectors can be mainly divided into two-stage methods [8,12,84] and one-stage methods [33,60,82,91,124] according to whether to use RoI-level operations [35,37]. Recently, Transformer-based detectors [9, 53, 131] have drawn great attention for their conceptually simple and flexible frameworks. Besides, instance segmentation approaches can also be divided into detector-based [8, 12, 37, 49, 90] and detector-free [16, 97] fashions according to whether box-level detectors are needed. Object detection and instance segmentation play critical roles and are foundations for all other instance perception tasks. For example, MOT, MOTS, and VIS extend image-level detection and segmentation to videos, requiring finding all object trajectories of specific classes in videos. Mainstream algorithms [47, 79, 101, 102, 109, 125] of MOT and MOTS follow an online "detection-then-association" paradigm. However, due to the intrinsic difference in benchmarks of MOTS [93, 120] (high-resolution long videos) and VIS [112] (low-resolution short videos), most recent VIS methods [43, 58, 100, 103] adopt an offline fashion. This strategy performs well on relatively simple VIS2019 [112], but the performance drops drastically on challenging OVIS [81] benchmark. Recently, IDOL [105] bridges the performance gap between online fashion and its offline counterparts by discriminative instance embeddings, showing the potential of the online paradigm in unifying MOT, MOTS, and VIS.

Retrieval by Language Expressions. REC, RES, and R-VOS aim at finding one specific target referred by a language expression using boxes or masks on the given images or videos. Similar to object detection, REC methods can be categorized into three paradigms: two-stage [40, 62, 65, 113], one-stage [57, 70, 114, 115], and Transformer-based [23, 45, 129] ones. Different from REC, RES approaches [11,25,31,41,44,69,119] focus more on designing diverse attention mechanisms to achieve vision-language alignment. Recently, SeqTR [130] unifies REC and RES as a point prediction problem and obtains promising results. Finally, R-VOS can be seen as a natural extension of RES from images to videos. Current state-of-the-art methods [7, 104] are Transformer-based and process the whole video in an offline fashion. However, the offline paradigm hinders the applications in the real world such as long videos and ongoing videos (e.g. autonomous driving).

Retrieval by Reference Annotations. SOT and VOS first specify tracked objects on the first frame of a video using boxes or masks, then require algorithms to predict the trajectories of the tracked objects in boxes or masks respectively. The core problems of these two tasks include (1) How to extract informative target features? (2) How to fuse the target information with representations of the current frame? For the first question, most SOT methods [5, 14, 50, 51, 110] encode target information by passing a template to a siamese backbone. While VOS approaches [18, 78, 117] usually pass multiple previous frames together with corresponding mask results to a memory encoder for extracting fine-grained target information. For the second question, correlations are widely adopted by early SOT algorithms [5, 51, 111]. However, these simple linear operations may cause serious information loss. To alleviate this problem, later works [14, 19, 110, 118] resort to Transformer for more discriminative representations. Besides, feature fusion in VOS is almost dominated by space-time memory networks [17, 18, 78, 117].

Unified Vision Models. Recently, unified vision models [13, 16, 34, 37, 54, 68, 83, 87, 95, 109, 132] have drawn great attention and achieved significant progress due to their strong generalizability and flexibility. Unified vision models attempt to solve multiple vision or multi-modal tasks by a single model. Existing works can be categorized into unified learning paradigms and unified model architectures.

Unified Learning Paradigms. These works [2, 34, 68, 83, 87, 95, 132] usually presents a universal learning paradigm for covering as many tasks and modalities as possible. For example, MuST [34] presents a multi-task self-training approach for 6 vision tasks. INTERN [87] introduces a continuous learning scheme, showing strong generalization ability on 26 popular benchmarks. Unified-IO [68] and OFA [95] proposes a unified sequence-to-sequence framework that can handle a variety of vision, language, and multi-modal tasks. Although these works can perform many tasks, the commonality and inner relationship among different tasks are less explored and exploited.

Unified Model Architectures. These works [13, 16, 37, 54, 109] usually designs a unified formulation or model architecture for a group of closely related tasks. For example, Mask R-CNN [37] proposes a unified network to perform object detection and instance segmentation simultaneously. Mask2Former [16] presents a universal architecture capable of handling panoptic, instance, and semantic segmentation. Pix2SeqV2 [13] designs a unified pixel-to-sequence interface for four vision tasks, namely object detection, instance segmentation, keypoint detection, and image captioning. GLIP [54] cleverly reformulates object detection as phrase grounding by replacing classical classification with word-region alignment. This new formulation allows joint training on both detection and grounding data, showing strong transferability to various object-level recognition tasks. However, GLIP [54] supports neither prompts in other modalities such as images & annotations nor video-level tracking tasks. In terms of object tracking, Unicorn [109] proposes a unified solution for SOT, VOS, MOT, and MOTS, achieving superior performance on 8 benchmarks with the same model weights. However, it is still difficult for Unicorn to handle diverse label vocabularies [22, 61, 74, 81, 112, 120] during training and inference. In this work, we propose a universal prompt-guided architecture for 10 instance perception tasks, conquering the drawbacks of GLIP [54] and Unicorn [109] simultaneously.

3. Approach

Before introducing detailed methods, we first categorize existing instance perception tasks into three classes.

- Object detection, instance segmentation, MOT, MOTS, and VIS take category names as prompts to find all instances of specific classes.
- REC, RES, and R-VOS exploit an expression as the prompt to localize a certain target.
- SOT and VOS use the annotation given in the first frame as the prompt for predicting the trajectories of the tracked target.

Essentially, all the above tasks aim to find objects specified by some prompts. This commonality motivates us to reformulate all instance perception tasks into a prompt-guided object discovery and retrieval problem and solve it by a unified model architecture and learning paradigm. As demonstrated in Figure 2, UNINEXT consists of three main components: (1) prompt generation (2) image-prompt feature fusion (3) object discovery and retrieval.

3.1. Prompt Generation

First, a prompt generation module is adopted to transform the original diverse prompt inputs into a unified form. According to different modalities, we introduce the corresponding strategies in the next two paragraphs respectively.

To deal with language-related prompts, a language encoder [24] $\mathrm{Enc_L}$ is adopted. To be specific, for category-guided tasks, we concatenate class names that appeared in the current dataset [61, 81, 112, 120] as the language expression. Take COCO [61] as an example, the expression can be written as "person. bicycle. ... . toothbrush". Then

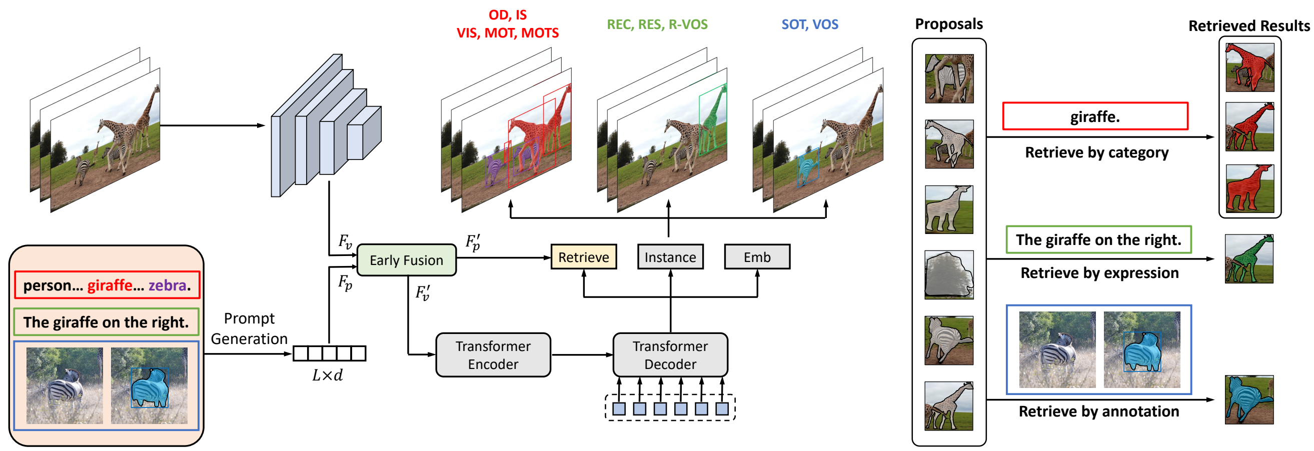

Figure 2. Framework of UNINEXT. The whole pipeline is shown on the left side. The schematic diagram of object retrieval is shown on the right side. The instance head predicts both boxes and masks of the objects. Better view in color on screen.

for both category-guided and expression-guided tasks, the language expression is passed into $\mathrm{Enc_L}$, getting a prompt embedding $F_p \in \mathbb{R}^{L\times d}$ with a sequence length of L.

For the annotation-guided tasks, to extract fine-grained visual features and fully exploit the target annotations, an additional reference visual encoder $\mathrm{Enc}_\mathrm{V}^\mathrm{ref}$ is introduced. Specifically, first a template with 2^2 times the target box area is cropped centered on the target location on the reference frame. Then the template is resized to a fixed size of 256×256. To introduce more precise target information, an extra channel named the target prior is concatenated to the template image, forming a 4-channel input. In more detail, the value of the target prior is 1 on the target region otherwise 0. Then the template image together with the target prior is passed to the reference visual encoder $\mathrm{Enc}_\mathrm{V}^\mathrm{ref}$, obtaining a hierarchical feature pyramid $\{\mathrm{C}_3, \mathrm{C}_4, \mathrm{C}_5, \mathrm{C}_6\}$. The corresponding spatial sizes are 32×32, 16×16, 8×8, and 4×4. To keep fine target information and get the prompt embedding in the same format as other tasks, a merging module is applied. Namely, all levels of features are first upsampled to 32×32 then added, and flattened as the final prompt embedding $F_p \in \mathbb{R}^{1024\times d}$.

The prompt generation process can be formulated as

$$F_p = \begin{cases} \mathrm{Enc}_\mathrm{L}^\mathrm{ref}(\text{expression}) & \text{expression-guided} \\ \mathrm{Enc}_\mathrm{L}^\mathrm{ref}(\text{concat}(\text{categories})) & \text{category-guided} \\ \text{merge}(\mathrm{Enc}_\mathrm{V}^\mathrm{ref}([\text{template, prior}]) & \text{annotation-guided} \end{cases}$$

3.2. Image-Prompt Feature Fusion

In parallel with the prompt generation, the whole current image is passed through another visual encoder $\mathrm{Enc_V}$, obtaining hierarchical visual features F_v. To enhance the original prompt embedding by the image contexts and to make the original visual features prompt-aware, an early fusion module is adopted. To be specific, first a bi-directional cross-attention module (Bi-XAtt) is used to retrieve information from different inputs, and then the retrieved representations are added to the original features. This process can be formulated as

$$\begin{aligned} F_\mathrm{p2v}, F_\mathrm{v2p} &= \text{Bi-XAtt}(F_v, F_p) \\ F'_v = F_v + F_\mathrm{p2v};\ F'_p &= F_p + F_\mathrm{v2p} \end{aligned} \tag{1}$$

Different from GLIP [54], which adopts 6 vision-language fusion layers and 6 additional BERT layers for feature enhancement, our early fusion module is much more efficient.

3.3. Object Discovery and Retrieval

With discriminative visual and prompt representations, the next crucial step is to transform input features into instances for various perception tasks. UNINEXT adopts the encoder-decoder architecture proposed by Deformable DETR [131] for its flexible query-to-instance fashion. We introduce the detailed architectures as follows.

The Transformer encoder takes hierarchical prompt-aware visual features as the inputs. With the help of efficient Multi-scale Deformable Self-Attention [131], target information from different scales can be fully exchanged, bringing stronger instance features for the subsequent instance decoding. Besides, as performed in two-stage Deformable DETR [131], an auxiliary prediction head is appended at the end of the encoder, generating N initial reference points with the highest scores as the inputs of the decoder.

The Transformer decoder takes the enhanced multi-scale features, N reference points from the encoder, as well as N object queries as the inputs. As shown in previous works [73, 100, 104, 123], object queries play a critical role in instance perception tasks. In this work, we attempt two query generation strategies: (1) static queries which do not change with images or prompts. (2) dynamic queries conditioned on the prompts. The first strategy can be easily implemented with `nn.Embedding(N,d)`. The second one can be performed by first pooling the enhanced prompt features F'_v along the sequence dimension, getting a global representation, then repeating it by N times. The above two methods are compared in Sec 4.3 and we find that static queries usually perform better than dynamic queries. The potential reason could be that static queries contain richer information and possess better training stability than dynamic queries. With the help of the deformable attention, the object queries can efficiently retrieve prompt-aware visual features and learn strong instance embedding $F_{\text{ins}} \in \mathbb{R}^{N\times d}$.

At the end of the decoder, a group of prediction heads is exploited to obtain the final instance predictions. Specifically, an instance head produces both boxes and masks of the targets. Besides, an embedding head [105] is introduced for associating the current detected results with previous trajectories in MOT, MOTS, and VIS. Until now, we have mined N potential instance proposals, which are represented with gray masks in Figure 2. However, not all proposals are what the prompts really refer to. Therefore, we need to further retrieve truly matched objects from these proposals according to the prompt embeddings as demonstrated in the right half of Figure 2. Specifically, given the prompt embeddings F'_p after early fusion, for category-guided tasks, we take the embedding of each category name as a weight matrix $W \in \mathbb{R}^{1\times d}$. Besides, for expression-guided and annotation-guided tasks, the weight matrix W is obtained by aggregating the prompt embedding F'_p using global average pooling (GAP) along the sequence dimension.

$$W = \begin{cases} F'_p[i],\ i \in \{0,1,...,C-1\} & \text{category} \\ \dfrac{1}{L}\displaystyle\sum_{i=0}^{L} F'_p(i,j) & \text{expression/annotation} \end{cases}$$

Finally, the instance-prompt matching scores S can be computed as the matrix multiplication of the target features and the transposed weight matrix. $S = F_{\text{ins}}W^{\top}$. Following previous work [54], the matching scores can be supervised by Focal Loss [60]. Different from previous fixed-size classifiers [131], the proposed retrieval head selects objects by the prompt-instance matching mechanism. This flexible design enables UNINEXT to jointly train on enormous datasets with diverse label vocabularies from different tasks, learning universal instance representations.

3.4. Training and Inference

Training. The whole training process consists of three consecutive stages: (1) general perception pretraining (2) image-level joint training (3) video-level joint training.

In the first stage, we pretrain UNINEXT on the large-scale object detection dataset Objects365 [88] for learning universal knowledge about objects. Since Objects365 does not have mask annotations, we introduce two auxiliary losses proposed by BoxInst [92] for training the mask branch. The loss function can be formulated as

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{retrieve}} + \mathcal{L}_{\text{box}} + \mathcal{L}^{\text{boxinst}}_{\text{mask}} \tag{2}$$

Then based on the pretrained weights of the first stage, we finetune UNINEXT jointly on image datasets, namely COCO [61] and the mixed dataset of RefCOCO [121], RefCOCO+ [121], and RefCOCOg [77]. With manually labeled mask annotations, the traditional loss functions like Dice Loss [75] and Focal Loss [60] can be used for the mask learning. After this step, UNINEXT can achieve superior performance on object detection, instance segmentation, REC, and RES.

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{retrieve}} + \mathcal{L}_{\text{box}} + \mathcal{L}_{\text{mask}} \tag{3}$$

Finally, we further finetune UNINEXT on video-level datasets for various downstream object tracking tasks and benchmarks. In this stage, the model is trained on two frames randomly chosen from the original videos. Besides, to avoid the model forgetting previously learned knowledge on image-level tasks, we also transform image-level datasets to pseudo videos for joint training with other video datasets. In summary, the training data in the third stage includes pseudo videos generated from COCO [61], RefCOCO/g/+ [77, 121, 121], SOT&VOS datasets (GOT-10K [42], LaSOT [28], TrackingNet [76], and Youtube-VOS [107]), MOT&VIS datasets (BDD100K [120], VIS19 [112], OVIS [81]), and R-VOS dataset Ref-Youtube-VOS [86]. Meanwhile, a reference visual encoder for SOT&VOS and an extra embedding head for association are introduced and optimized in this period.

$$\mathcal{L}_{\text{stage3}} = \mathcal{L}_{\text{retrieve}} + \mathcal{L}_{\text{box}} + \mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{embed}} \tag{4}$$

Inference. For category-guided tasks, UNINEXT predicts instances of different categories and associates them with previous trajectories. The association proceeds in an online fashion and is purely based on the learned instance embedding following [79, 105]. For expression-guided and annotation-guided tasks, we directly pick the object with the highest matching score with the given prompt as the final result. Different from previous works [94, 104] restricted by the offline fashion or complex post-processing, our method is simple, online, and post-processing free.

4. Experiments

4.1. Implementation Details

We attempt two different backbones, ResNet-50 [38] and ConvNeXt-Large [66] as the visual encoder. We adopt BERT [24] as the text encoder and its parameters are trained in the first and second training stages while being frozen in

Table 1. State-of-the-art comparison on object detection.

Model	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster R-CNN [84]	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 [89]		45.0	63.4	48.2	26.9	47.2	59.5
Cascade Mask-RCNN [8]		46.3	64.3	50.5	-	-	-
Deformable-DETR [131]		46.9	65.6	51.0	29.6	50.1	61.6
DN-Deformable-DETR [53]		<u>48.6</u>	<u>67.4</u>	<u>52.7</u>	<u>31.0</u>	<u>52.0</u>	<u>63.7</u>
UNINEXT		**51.3**	**68.4**	**56.2**	**32.6**	**55.7**	**66.5**
Cascade Mask R-CNN [8]	ConvNeXt-L	<u>54.8</u>	<u>73.8</u>	<u>59.8</u>	-	-	-
UNINEXT		**58.1**	**74.9**	**63.7**	**40.7**	**62.5**	**73.6**

Table 2. State-of-the-art comparison on instance segmentation.

Model	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
CondInst [90]	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 [99]		38.8	59.9	41.7	16.5	41.7	<u>56.2</u>
HTC [12]		39.7	61.4	43.1	22.6	42.2	50.6
RefineMask [122]		40.2	-	-	-	-	-
QueryInst [29]		<u>40.6</u>	<u>63.0</u>	<u>44.0</u>	<u>23.4</u>	<u>42.5</u>	52.8
UNINEXT		**44.9**	**67.0**	**48.9**	**26.3**	**48.5**	**59.0**
Cascade Mask R-CNN [8]	Swin-L	46.7	70.1	50.8	-	-	-
QueryInst [29]		<u>49.1</u>	**74.2**	<u>53.8</u>	**31.5**	<u>51.8</u>	<u>63.2</u>
Cascade Mask R-CNN [8]	ConvNeXt-L	47.6	71.3	51.7	-	-	-
UNINEXT		**49.6**	<u>73.4</u>	**54.3**	<u>30.4</u>	**53.6**	**65.7**

Table 3. State-of-the-art comparison on REC.

Method	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val-u	test-u
UNITER_L [15]	81.41	87.04	74.17	75.90	81.45	66.70	74.86	75.77
VILLA_L [32]	82.39	87.48	74.84	76.17	81.54	66.84	76.18	76.71
MDETR [45]	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89
RefTR [55]	85.65	88.73	81.16	77.55	82.26	68.99	79.25	80.01
SeqTR [130]	87.00	90.15	83.59	78.69	84.51	71.87	82.69	83.37
UNINEXT-R50	<u>89.72</u>	<u>91.52</u>	<u>86.93</u>	<u>79.76</u>	<u>85.23</u>	<u>72.78</u>	<u>83.95</u>	<u>84.31</u>
UNINEXT-L	**91.43**	**93.73**	**88.93**	**83.09**	**87.90**	**76.15**	**86.91**	**87.48**

Table 4. State-of-the-art comparison on RES.

Method	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val-u	test-u
CMSA [119]	58.32	60.61	55.09	43.76	47.60	37.89	-	-
BRINet [41]	60.98	62.99	59.21	48.17	52.32	42.11	-	-
CMPC+ [63]	62.47	65.08	60.82	50.25	54.04	43.47	-	-
MCN [70]	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40
EFN [31]	62.76	65.69	59.67	51.50	55.24	43.01	-	-
VLT [25]	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65
SeqTR [130]	71.70	73.31	69.82	63.04	66.73	58.97	64.69	65.74
LAVT [116]	72.73	75.82	68.79	62.14	68.38	55.10	61.24	62.09
UNINEXT-R50	<u>77.90</u>	<u>79.68</u>	<u>75.77</u>	<u>66.20</u>	<u>71.22</u>	<u>59.01</u>	<u>70.04</u>	<u>70.52</u>
UNINEXT-L	**80.32**	**82.61**	**77.76**	**70.04**	**74.91**	**62.57**	**73.41**	**73.68**

the last training stage. The Transformer encoder-decoder architecture follows [131] with 6 encoder layers and 6 decoder layers. The number of object queries N is set to 900. The optimizer is AdamW [67] with weight decay of 0.05. The model is trained on 32 and 16 A100 GPUs for Objects365 pretraining and other stages respectively. More details can be found in the supplementary materials.

4.2. Evaluations on 10 Tasks

We compare UNINEXT with task-specific counterparts in 20 datasets. In each benchmark, the best two results are indicated in **bold** and with <u>underline</u>. UNINEXT in all benchmarks uses the same model parameters.

Object Detection and Instance Segmentation. We compare UNINEXT with state-of-the-art object detection and instance segmentation methods on COCO `val2017` ($5k$ images) and `test-dev` split ($20k$ images) respectively. As shown in Table 1, UNINEXT surpasses state-of-the-art query-based detector DN-Deformable DETR [53] by 2.7 AP. By replacing ResNet-50 [38] with stronger ConvNeXt-Large [66] as the backbone, UNINEXT achieves a box AP of 58.1, surpassing competitive rivals Cascade Mask-RCNN [8] by 3.3. Besides, the results of instance segmentation are shown in Table 2. With the same ResNet-50 backbone, UNINEXT outperforms state-of-the-art QueryInst by 4.3 AP and 6.2 AP_L. When using ConvNeXt-Large as the backbone, UNINEXT achieves a mask AP of 49.6, surpassing Cascade Mask R-CNN [8] by 2.0.

REC and RES. RefCOCO [121], RefCOCO+ [121], and RefCOCOg [71] are three representative benchmarks for REC and RES proposed by different institutions. Following previous literature, we adopt Precision@0.5 and overall intersection-over-union (oIoU) as the evaluation metrics for REC and RES respectively and results are rounded to two decimal places. As shown in Table 3 and Table 4, our method with ResNet-50 backbone surpasses all previous approaches on all splits. Furthermore, when using ConvNeXt-Large backbone, UNINEXT obtains new state-of-the-art results, exceeding the previous best method by a large margin. Especially on RES, UNINEXT-L outperforms LAVT [116] by 8.63 on average.

SOT. We compare UNINEXT with state-of-the-art SOT methods on four large-scale benchmarks: LaSOT [28], LaSOT-ext [27], TrackingNet [76], and TNL-2K [98]. These benchmarks adopt the area under the success curve (AUC), normalized precision (P_{Norm}), and precision (P) as the evaluation metrics and include 280, 150, 511, and 700 videos in the test set respectively. As shown in Table 5, UNINEXT achieves the best results in terms of AUC and P among all trackers with ResNet-50 backbone. Especially on TNL-2K, UNINEXT outperforms the second best method TransT [14] by 5.3 AUC and 5.8 P respectively. Besides, UNINEXT with ConvNeXt-Large backbone obtains the best AUC on all four benchmarks, exceeding Unicorn [109] with the same backbone by 3.9 on LaSOT.

VOS. The comparisons between UNINEXT with previous semi-supervised VOS methods are demonstrated in Table 6. DAVIS-2017 [80] adopts region similarity $\mathcal{J}$, contour accuracy $\mathcal{F}$, and the averaged score $\mathcal{J}\&\mathcal{F}$ as the metrics. Similarly, Youtube-VOS 2018 [107] reports $\mathcal{J}$ and $\mathcal{F}$ for both seen and unseen categories, and the averaged overall score $\mathcal{G}$. UNINEXT achieves the best results among all non-memory-based methods, further bridging the per-

Table 5. State-of-the-art comparison on SOT.

Method	Backbone	LaSOT [28]			LaSOT$_{\text{ext}}$ [27]			TrackingNet [76]			TNL-2K [98]	
		AUC	P_{Norm}	P	AUC	P_{Norm}	P	AUC	P_{Norm}	P	AUC	P
PrDiMP [21]	ResNet-50	59.8	68.8	60.8	-	-	-	75.8	81.6	70.4	47.0	45.9
LTMU [20]	ResNet-50	57.2	-	57.2	41.4	49.9	47.3	-	-	-	48.5	47.3
TransT [14]	ResNet-50	64.9	73.8	69.0	-	-	-	81.4	86.7	80.3	50.7	51.7
KeepTrack [72]	ResNet-50	67.1	**77.2**	70.2	48.2	-	-	-	-	-	-	-
UNINEXT	ResNet-50	**69.2**	77.1	**75.5**	**51.2**	**58.1**	**58.1**	**83.2**	**86.9**	**83.3**	**56.0**	**57.5**
SimTrack [10]	ViT-B	69.3	78.5	-	-	-	-	82.3	-	**86.5**	54.8	53.8
OSTrack [118]	ViT-B	71.1	**81.1**	77.6	50.5	61.3	57.6	83.9	**88.5**	83.2	55.9	**-**
Unicorn [109]	ConvNeXt-L	68.5	76.6	74.1	-	-	-	83.0	86.4	82.2	-	-
UNINEXT	ConvNeXt-L	**72.4**	80.7	**78.9**	**54.4**	**61.8**	**61.4**	**85.1**	88.2	84.7	**58.1**	**60.7**

Table 6. State-of-the-art comparison on VOS.

	Method	YT-VOS 2018 val [107]					DAVIS 2017 val [80]		
		$\mathcal{G}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Memory	STM [78]	79.4	79.7	84.2	72.8	80.9	81.8	79.2	84.3
Memory	CFBI [117]	81.4	81.1	85.8	75.3	83.4	81.9	79.1	84.6
Memory	STCN [18]	83.0	81.9	86.5	77.9	85.7	85.4	82.2	88.6
Memory	XMem [17]	86.1	85.1	89.8	80.3	89.2	87.7	84.0	91.4
Non-Memory	SiamMask [96]	52.8	60.2	58.2	45.1	47.7	56.4	54.3	58.5
Non-Memory	Unicorn [109]	-	-	-	-	-	69.2	65.2	73.2
Non-Memory	Siam R-CNN [94]	73.2	73.5	-	66.2	-	70.6	66.1	75.0
Non-Memory	TVOS [127]	67.8	67.1	69.4	63.0	71.6	72.3	69.9	74.7
Non-Memory	FRTM [85]	72.1	72.3	76.2	65.9	74.1	76.7	**73.9**	79.6
Non-Memory	**UNINEXT-R50**	77.0	76.8	81.0	70.8	79.4	74.5	71.3	77.6
Non-Memory	**UNINEXT-L**	**78.1**	**79.1**	**83.5**	**71.0**	**78.9**	**77.2**	73.2	**81.2**

Table 7. State-of-the-art comparison on MOT.

Method	Split	mMOTA↑	mIDF1↑	MOTA↑	IDF1↑	ID Sw.↓
Yu *et al.* [120]	val	25.9	44.5	56.9	66.8	8315
QDTrack [79]	val	36.6	50.8	63.5	**71.5**	**6262**
Unicorn [109]	val	41.2	54.0	**66.6**	71.3	10876
UNINEXT-L	val	**41.8**	**54.9**	64.6	68.7	9134

Table 8. State-of-the-art comparison on MOTS.

Method	Online	mMOTSA↑	mMOTSP↑	mIDF1↑	ID Sw.↓
MaskTrackRCNN [112]	✓	12.3	59.9	26.2	9116
STEm-Seg [1]	✗	12.2	58.2	25.4	8732
QDTrack-mots [79]	✓	22.5	59.6	40.8	1340
PCAN [47]	✓	27.4	66.7	45.1	876
VMT [46]	✗	28.7	67.3	**45.7**	**825**
Unicorn [109]	✓	29.6	**67.7**	44.2	1731
UNINEXT-L	✓	**32.0**	60.2	45.4	1634

formance gap between non-memory-based approaches and memory-based ones. Furthermore, compared with traditional memory-based methods [18,78], UNINEXT does not rely on the intermediate mask predictions. This leads to constant memory consumption and zero cumulative error, enabling UNINEXT to handle long sequences of any length.

MOT. We compare UNINEXT with state-of-the-art MOT methods on BDD100K [120], which requires tracking 8 classes of instances in the autonomous driving scenario. Except for classical evaluation metrics Multiple-Object Tracking Accuracy (MOTA), Identity F1 Score (IDF1), and Identity Switches (IDS), BDD100K additionally introduces mMOTA, and mIDF1 to evaluate the average performance across 8 classes. As shown in Table 7, UNINEXT surpasses Unicorn [109] by 0.6 mMOTA and 0.9 mIDF1 respectively.

MOTS. Similar to MOT, BDD100K MOTS Challenge [120] evaluates the performance on multi-class tracking by mMOTSA, mMOTSP, mIDF1, and ID Sw. This benchmark contains 37 sequences with mask annotations in the validation set. As shown in Table 8, UNINEXT achieves state-of-the-art performance, surpassing the previous best method Unicorn [109] by 2.4 mMOTSA.

VIS. We compare UNINEXT against state-of-the-art VIS methods on Youtube-VIS 2019 [112] and OVIS [81] validation sets. Specifically, Youtube-VIS 2019 and OVIS have 40 and 25 object categories, containing 302 and 140 videos respectively in the validation set. Both benchmarks take AP as the main metric. When using the same ResNet-50 backbone, UNINEXT obtains the best results on both datasets. Especially on more challenging OVIS, UNINEXT exceeds the previous best method IDOL [105] by 3.8 AP. As shown in Table 9, when using stronger ConvNeXt-Large backbone, UNINEXT achieves much higher AP (64.3 on Youtube-VIS 2019 and 41.1 on OVIS). Although the performance is slightly lower than IDOL, UNINEXT can handle diverse label vocabularies [61, 81, 112, 120] by a single model, being much more flexible than IDOL.

R-VOS. Ref-Youtube-VOS [86] and Ref-DAVIS17 [48] are two popular R-VOS benchmarks, which are constructed by introducing language expressions for the objects in the original Youtube-VOS [107] and DAVIS17 [80] datasets. As same as semi-supervised VOS, region similarity $\mathcal{J}$, contour accuracy $\mathcal{F}$, and the averaged score $\mathcal{J}\&\mathcal{F}$ are adopted as the metrics. As demonstrated in Table 10, UNINEXT outperforms all previous R-VOS approaches by a large margin, when using the same ResNet-50 backbone. Especially on Ref-DAVIS17, UNINEXT exceeds previous best ReferFormer [104] by 5.4 $\mathcal{J}\&\mathcal{F}$. Furthermore, when adopting stronger ConvNeXt-Large backbone, UNINEXT achieves new state-of-the-art $\mathcal{J}\&\mathcal{F}$(66.2 on Ref-Youtube-VOS and 66.7 on Ref-DAVIS17). Besides, different from offline Ref-

Table 9. State-of-the-art comparison on VIS.

Method	Backbone	Online	VIS2019 val			OVIS val		
			AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
VisTR [100]	ResNet-50	✗	36.2	59.8	36.9	-	-	-
MaskProp [4]		✗	40.0	-	42.9	-	-	-
IFC [43]		✗	42.8	65.8	46.8	13.1	27.8	11.6
SeqFormer [103]		✗	47.4	69.8	51.8	15.1	31.9	13.8
IDOL [105]		✓	49.5	74.0	52.9	30.2	51.3	30.0
VITA [39]		✗	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 [103]	Swin-L	✗	59.3	82.1	66.4	-	-	-
VMT [46]		✗	59.7	-	66.7	19.8	39.6	17.2
VITA [39]		✗	63.0	86.9	67.9	-	-	-
IDOL [105]		✓	**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

Table 10. State-of-the-art comparison on 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 [119]	ResNet-50	36.4	34.8	38.1	40.2	36.9	43.5
URVOS [86]		47.2	45.3	49.2	51.5	47.3	56.0
YOFO [52]		48.6	47.5	49.7	54.4	50.1	58.7
ReferFormer [104]		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 [26]	Ensemble	54.2	53.0	55.5	-	-	-
CITD [56]		61.4	60.0	62.7	-	-	-
MTTR [7]	Video-Swin-T	55.3	54.0	56.6	-	-	-
ReferFormer [104]		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**

Table 11. Ablations. The settings in our final model is underlined.

Experiment	Method	OD COCO (AP)	REC RefCOCO (P@0.5)	VOS YTBVOS ($\mathcal{J}\&\mathcal{F}$)	RVOS R-YTBVOS ($\mathcal{J}\&\mathcal{F}$)	VIS VIS19 (AP)
Fusion	Early Fusion	51.3	89.7	77.0	61.2	53.0
	W/o Fusion	51.1	87.4	55.6	58.4	51.0
		(+0.2)	(+2.3)	(+21.4)	(+2.8)	(+2.0)
Queries	Static	51.3	89.7	77.0	61.2	53.0
	Dynamic	51.9	89.8	77.4	61.6	50.2
		(-0.6)	(-0.1)	(-0.4)	(-0.4)	(+2.8)
Model	Unified	51.3	89.7	77.0	61.2	53.0
	Task-specific	50.8	87.6	74.2	57.2	50.1
		(+0.5)	(+2.1)	(+2.8)	(+4.0)	(+2.9)

Former, UNINEXT works in a flexible online fashion, making it applicable to ongoing videos in the real world.

4.3. Ablations and Other Analysis

In this section, we conduct component-wise analysis for better understanding our method. All models take ResNet-50 as the backbone. The methods are evaluated on five benchmarks (COCO [61], RefCOCO [121], Youtube-VOS [107], Ref-Youtube-VOS [86], and Youtube-VIS 2019 [112]) from five tasks (object detection, REC, VOS, R-VOS, and VIS). The results are shown in Table 11.

Fusion. To study the effect of feature fusion between visual features and prompt embeddings, we implement a variant without any early fusion. In this version, prompt embeddings do not have an influence on proposal generation but are only used in the final object retrieval process. Experiments show that early fusion has the greatest impact on VOS, the performance on VOS drops drastically by 21.4 $\mathcal{J}\&\mathcal{F}$ without feature fusion. This is mainly caused by the following reasons (1) Without the guidance of prompt embeddings, the network can hardly find rare referred targets like trees and sinks. (2) Without early fusion, the network cannot fully exploit fine mask annotations in the first frame, causing degradation of the mask quality. Besides, the removal of feature fusion also causes performance drop of 2.3 P@0.5 and 2.8 $\mathcal{J}\&\mathcal{F}$on REC and RVOS respectively, showing the importance of early fusion in expression-guided tasks. Finally, feature fusion has minimum influence on object detection and VIS. This can be understood because both two tasks aim to find all objects as completely as possible rather than locating one specific target referred by the prompt.

Queries. We compare two different query generation strategies: static queries by `nn.Embedding(N, d)` and dynamic queries conditioned on the prompt embeddings. Experiments show that dynamic queries perform slightly better than static queries on the first four tasks. However, static queries outperform dynamic ones by 2.8 AP on the VIS task, obtaining higher overall performance. A potential reason is that N different object queries can encode richer inner relationship among different targets than simply copying the pooled prompt by N times as queries. This is especially important for VIS because targets need to be associated according to their affinity in appearance and space.

Unification. We compare two different model design philosophies, one unified model or multiple task-specific models. Except for the unified model, we also retrain five task-specific models only on data from corresponding tasks. Experiments show that the unified model achieves better performance than its task-specific counterparts on five tasks, demonstrating the superiority of the unified formulation and joint training on all instance perception tasks. Finally, the unified model can save tons of parameters, being much more parameter-efficient.

5. Conclusions

We propose UNINEXT, a universal instance perception model of the next generation. For the first time, UNINEXT unifies 10 instance perception tasks with a prompt-guided object discovery and retrieval paradigm. Extensive experiments demonstrate that UNINEXT achieves superior performance on 20 challenging benchmarks with a single model with the same model parameters. We hope that UNINEXT can serve as a solid baseline for the research of instance perception in the future.

Acknowledgement. We would like to thank the reviewers for their insightful comments. The paper is supported in part by the National Key R&D Program of China under Grant No. 2018AAA0102001, 2022ZD0161000 and National Natural Science Foundation of China under grant No. 62293542, U1903215, 62022021.

References

[1] Ali Athar, Sabarinath Mahadevan, Aljosa Osep, Laura Leal-Taixé, and Bastian Leibe. STEm-Seg: Spatio-temporal embeddings for instance segmentation in videos. In *ECCV*, 2020. 7

[2] Paul Barham, Aakanksha Chowdhery, Jeff Dean, Sanjay Ghemawat, Steven Hand, Daniel Hurt, Michael Isard, Hyeontaek Lim, Ruoming Pang, Sudip Roy, et al. Pathways: Asynchronous distributed dataflow for ML. *PMLS*, 2022. 3

[3] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *ICCV*, 2019. 1

[4] Gedas Bertasius and Lorenzo Torresani. Classifying, segmenting, and tracking object instances in video with mask propagation. In *CVPR*, 2020. 8

[5] Luca Bertinetto, Jack Valmadre, João F Henriques, Andrea Vedaldi, and Philip H S Torr. Fully-convolutional siamese networks for object tracking. In *ECCVW*, 2016. 1, 3

[6] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. YOLACT: Real-time instance segmentation. In *ICCV*, 2019. 1

[7] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multi-modal transformers. In *CVPR*, 2022. 1, 3, 8

[8] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: high quality object detection and instance segmentation. *TPAMI*, 2019. 1, 2, 6

[9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 1, 2, 6

[10] Boyu Chen, Peixia Li, Lei Bai, Lei Qiao, Qiuhong Shen, Bo Li, Weihao Gan, Wei Wu, and Wanli Ouyang. Backbone is all your need: A simplified architecture for visual object tracking. In *ECCV*, 2022. 7

[11] Ding-Jie Chen, Songhao Jia, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. See-through-text grouping for referring image segmentation. In *ICCV*, 2019. 2

[12] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *CVPR*, 2019. 2, 6

[13] Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J Fleet, and Geoffrey Hinton. A unified sequence interface for vision tasks. *NeurIPS*, 2022. 3

[14] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *CVPR*, 2021. 3, 6, 7

[15] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 6

[16] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 2, 3

[17] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *ECCV*, 2022. 3, 7

[18] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *NeurIPS*, 2021. 1, 3, 7

[19] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *CVPR*, 2022. 3

[20] Kenan Dai, Yunhua Zhang, Dong Wang, Jianhua Li, Huchuan Lu, and Xiaoyun Yang. High-performance long-term tracking with meta-updater. In *CVPR*, 2020. 7

[21] Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic regression for visual tracking. In *CVPR*, 2020. 7

[22] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. TAO: A large-scale benchmark for tracking any object. In *ECCV*, 2020. 2, 3

[23] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. TransVG: End-to-end visual grounding with transformers. In *ICCV*, 2021. 2

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*, 2019. 3, 5

[25] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *ICCV*, 2021. 2, 6

[26] Zihan Ding, Tianrui Hui, Shaofei Huang, Si Liu, Xuan Luo, Junshi Huang, and Xiaoming Wei. Progressive multimodal interaction network for referring video object segmentation. *The 3rd Large-scale Video Object Segmentation Challenge*, 2021. 8

[27] Heng Fan, Hexin Bai, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Mingzhen Huang, Juehuan Liu, Yong Xu, et al. LaSOT: A high-quality large-scale single object tracking benchmark. *IJCV*, 2021. 6, 7

[28] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. LaSOT: A high-quality benchmark for large-scale single object tracking. In *CVPR*, 2019. 5, 6, 7

[29] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries. In *ICCV*, 2021. 6

[30] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 2010. 1

[31] Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. Encoder fusion network with co-attention embedding for referring image segmentation. In *CVPR*, 2021. 2, 6

[32] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *NeurIPS*, 2020. 6

[33] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 2

[34] Golnaz Ghiasi, Barret Zoph, Ekin D Cubuk, Quoc V Le, and Tsung-Yi Lin. Multi-task self-training for learning general representations. In *ICCV*, 2021. 3

[35] Ross Girshick. Fast R-CNN. In *ICCV*, 2015. 2

[36] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 2

[37] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 1, 2, 3

[38] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5, 6

[39] Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. Vita: Video instance segmentation via object token association. *NeurIPS*, 2022. 8

[40] Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. Learning to compose and reason with language tree structures for visual grounding. *TPAMI*, 2019. 2

[41] Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. Bi-directional relationship inferring network for referring image segmentation. In *CVPR*, 2020. 2, 6

[42] Lianghua Huang, Xin Zhao, and Kaiqi Huang. GOT-10k: A large high-diversity benchmark for generic object tracking in the wild. *TPAMI*, 2019. 5

[43] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Video instance segmentation using inter-frame communication transformers. *NeurIPS*, 2021. 1, 2, 8

[44] Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li, and Tieniu Tan. Locate then segment: A strong pipeline for referring image segmentation. In *CVPR*, 2021. 2

[45] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*, 2021. 2, 6

[46] Lei Ke, Henghui Ding, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Video mask transfiner for high-quality video instance segmentation. *ECCV*, 2022. 7, 8

[47] Lei Ke, Xia Li, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Prototypical cross-attention networks for multiple object tracking and segmentation. *NeurIPS*, 2021. 1, 2, 7

[48] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *ACCV*, 2018. 7

[49] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *CVPR*, 2020. 2

[50] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. SiamRPN++: Evolution of siamese visual tracking with very deep networks. In *CVPR*, 2019. 3

[51] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *CVPR*, 2018. 1, 3

[52] Dezhuang Li, Ruoqi Li, Lijun Wang, Yifan Wang, Jinqing Qi, Lu Zhang, Ting Liu, Qingquan Xu, and Huchuan Lu. You only infer once: Cross-modal meta-transfer for referring video object segmentation. In *AAAI*, 2022. 8

[53] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *CVPR*, 2022. 2, 6

[54] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, 2022. 3, 4, 5

[55] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *NeurIPS*, 2021. 6

[56] Chen Liang, Yu Wu, Tianfei Zhou, Wenguan Wang, Zongxin Yang, Yunchao Wei, and Yi Yang. Rethinking cross-modal interaction from a top-down perspective for referring video object segmentation. *arXiv preprint arXiv:2106.01061*, 2021. 8

[57] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A real-time cross-modality correlation filtering method for referring expression comprehension. In *CVPR*, 2020. 2

[58] Huaijia Lin, Ruizheng Wu, Shu Liu, Jiangbo Lu, and Jiaya Jia. Video instance segmentation with a propose-reduce paradigm. In *ICCV*, 2021. 2

[59] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 1

[60] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 2, 5

[61] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 2, 3, 5, 7, 8

[62] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *ICCV*, 2019. 2

[63] Si Liu, Tianrui Hui, Shaofei Huang, Yunchao Wei, Bo Li, and Guanbin Li. Cross-modal progressive comprehension for referring segmentation. *TPAMI*, 2021. 6

[64] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *CVPR*, 2018. 1

[65] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. Improving referring expression grounding with cross-modal attention-guided erasing. In *CVPR*, 2019. 2

[66] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *CVPR*, 2022. 5, 6

[67] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[68] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-IO: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022. 3

[69] Gen Luo, Yiyi Zhou, Rongrong Ji, Xiaoshuai Sun, Jinsong Su, Chia-Wen Lin, and Qi Tian. Cascade grouped

attention network for referring expression segmentation. In *ACMMM*, 2020. 2

[70] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *CVPR*, 2020. 2, 6

[71] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 6

[72] Christoph Mayer, Martin Danelljan, Danda Pani Paudel, and Luc Van Gool. Learning target candidate association to keep track of what not to track. In *ICCV*, 2021. 7

[73] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *CVPR*, 2022. 4

[74] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. MOT16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 1, 2, 3

[75] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 2016. 5

[76] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *ECCV*, 2018. 5, 6, 7

[77] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *ECCV*, 2016. 5

[78] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, 2019. 1, 3, 7

[79] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking. In *CVPR*, 2021. 2, 5, 7

[80] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 Davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 6, 7

[81] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *IJCV*, 2022. 2, 3, 5, 7

[82] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2015. 1, 2

[83] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022. 3

[84] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R–CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1, 2, 6

[85] Andreas Robinson, Felix Jaremo Lawin, Martin Danelljan, Fahad Shahbaz Khan, and Michael Felsberg. Learning fast and robust target models for video object segmentation. In *CVPR*, 2020. 7

[86] Seonguk Seo, Joon-Young Lee, and Bohyung Han. URVOS: Unified referring video object segmentation network with a large-scale benchmark. In *ECCV*, 2020. 1, 5, 7, 8

[87] Jing Shao, Siyu Chen, Yangguang Li, Kun Wang, Zhenfei Yin, Yinan He, Jianing Teng, Qinghong Sun, Mengya Gao, Jihao Liu, et al. INTERN: A new learning paradigm towards general vision. *arXiv preprint arXiv:2111.08687*, 2021. 3

[88] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, 2019. 2, 5

[89] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse R-CNN: End-to-end object detection with learnable proposals. In *CVPR*, 2021. 6

[90] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *ECCV*, 2020. 1, 2, 6

[91] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *ICCV*, 2019. 1, 2

[92] Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. BoxInst: High-performance instance segmentation with box annotations. In *CVPR*, 2021. 5

[93] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. MOTS: Multi-object tracking and segmentation. In *CVPR*, 2019. 1, 2

[94] Paul Voigtlaender, Jonathon Luiten, Philip HS Torr, and Bastian Leibe. Siam R-CNN: Visual tracking by re-detection. In *CVPR*, 2020. 5, 7

[95] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, 2022. 3

[96] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip H. S. Torr. Fast online object tracking and segmentation: A unifying approach. In *CVPR*, 2019. 7

[97] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. SOLO: Segmenting objects by locations. In *ECCV*, 2020. 1, 2

[98] Xiao Wang, Xiujun Shu, Zhipeng Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In *CVPR*, 2021. 6, 7

[99] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. SOLOv2: Dynamic and fast instance segmentation. *NeurIPS*, 2020. 6

[100] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *CVPR*, 2021. 1, 2, 4, 8

[101] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking. In *ECCV*, 2020. 2

[102] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, 2017. 2

[103] J Wu, Y Jiang, S Bai, W Zhang, and X Bai. Seqformer: Sequential transformer for video instance segmentation. *ECCV*, 2021. 1, 2, 8

[104] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *CVPR*, 2022. 1, 3, 4, 5, 7, 8

[105] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. *ECCV*, 2022. 2, 5, 7, 8

[106] Yi Wu, Jongwoo Lim, and Ming Hsuan Yang. Object tracking benchmark. *TPAMI*, 2015. 1

[107] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. YouTube-VOS: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 1, 5, 6, 7, 8

[108] Zhenbo Xu, Wei Yang, Wei Zhang, Xiao Tan, Huan Huang, and Liusheng Huang. Segment as points for efficient and effective online multi-object tracking and segmentation. *TPAMI*, 2021. 1

[109] Bin Yan, Yi Jiang, Peize Sun, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Towards grand unification of object tracking. In *ECCV*, 2022. 2, 3, 6, 7

[110] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *ICCV*, 2021. 3

[111] Bin Yan, Xinyu Zhang, Dong Wang, Huchuan Lu, and Xiaoyun Yang. Alpha-refine: Boosting tracking performance by precise bounding box estimation. In *CVPR*, 2021. 3

[112] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019. 1, 2, 3, 5, 7, 8

[113] Sibei Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *ICCV*, 2019. 2

[114] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-query construction. In *ECCV*, 2020. 2

[115] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *ICCV*, 2019. 1, 2

[116] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. LAVT: Language-aware vision transformer for referring image segmentation. In *CVPR*, 2022. 1, 6

[117] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *ECCV*, 2020. 3, 7

[118] Botao Ye, Hong Chang, Bingpeng Ma, and Shiguang Shan. Joint feature learning and relation modeling for tracking: A one-stream framework. In *ECCV*, 2022. 3, 7

[119] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *CVPR*, 2019. 1, 2, 6, 8

[120] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020. 2, 3, 5, 7

[121] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016. 1, 5, 6, 8

[122] Gang Zhang, Xin Lu, Jingru Tan, Jianmin Li, Zhaoxiang Zhang, Quanquan Li, and Xiaolin Hu. RefineMask: Towards high-quality instance segmentation with fine-grained features. In *CVPR*, 2021. 6

[123] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. DINO: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 4

[124] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *CVPR*, 2020. 2

[125] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *ECCV*, 2022. 2

[126] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. FairMOT: On the fairness of detection and re-identification in multiple object tracking. *IJCV*, 2021. 1

[127] Yizhuo Zhang, Zhirong Wu, Houwen Peng, and Stephen Lin. A transductive approach for video object segmentation. In *CVPR*, 2020. 7

[128] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *ECCV*, 2020. 1

[129] Yiyi Zhou, Tianhe Ren, Chaoyang Zhu, Xiaoshuai Sun, Jianzhuang Liu, Xinghao Ding, Mingliang Xu, and Rongrong Ji. TRAR: Routing the attention spans in transformer for visual question answering. In *ICCV*, 2021. 2

[130] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. SeqTR: A simple yet universal network for visual grounding. *ECCV*, 2022. 1, 2, 6

[131] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2020. 2, 4, 5, 6

[132] Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *CVPR*, 2022. 3