Generalized Decoding for Pixel, Image, and Language

Xueyan Zou*§, Zi-Yi Dou*♯, Jianwei Yang*‡♠, Zhe Gan†, Linjie Li†, Chunyuan Li‡, Xiyang Dai†, Harkirat Behl‡
Jianfeng Wang†, Lu Yuan†, Nanyun Peng♯, Lijuan Wang†, Yong Jae Lee¶§, Jianfeng Gao¶‡
§ University of Wisconsin-Madison ♯ UCLA ‡ Microsoft Research at Redmond † Microsoft Cloud & AI
*Equal Technical Contribution ¶ Equal Advisory Contribution ♠ Project Lead
{xueyan,yongjaelee}@cs.wisc.edu {zdou,violetpeng}@cs.ucla.edu {jianwyan,jfgao,zhgan,linjli,chunyl,jianfw,luyuan,lijuanw,hbehl,xidai}@microsoft.com

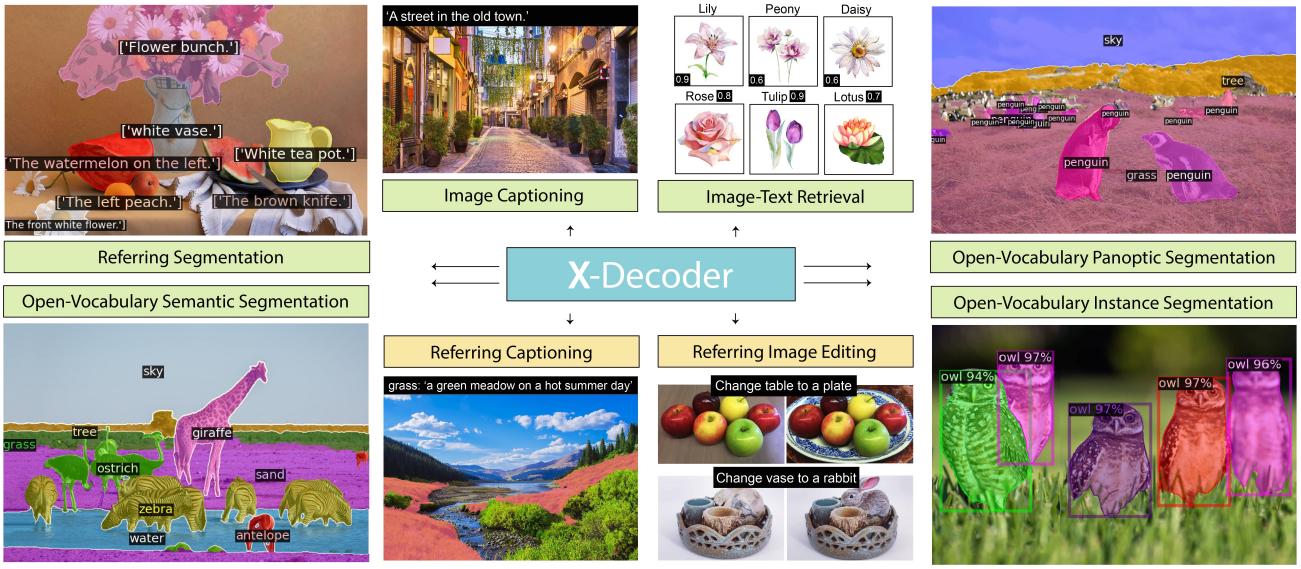

Figure 1. With one suite of parameters, X-Decoder after pretraining supports all types of image segmentation tasks ranging from open-vocabulary instance/semantic/panoptic segmentation to referring segmentation, and vision-language tasks including image-text retrieval, and image captioning (labeled in green boxes). It further empowers composite tasks like referring captioning using X-Decoder itself and image editing collaborating with generative models such as Stable Diffusion [61] (labeled in yellow boxes).

Abstract

We present X-Decoder, a generalized decoding model that can predict pixel-level segmentation and language tokens seamlessly. X-Decoder takes as input two types of queries: (i) generic non-semantic queries and (ii) semantic queries induced from text inputs, to decode different pixel-level and token-level outputs in the same semantic space. With such a novel design, X-Decoder is the first work that provides a unified way to support all types of image segmentation and a variety of vision-language (VL) tasks. Without any pseudo-labeling, our design enables seamless interactions across tasks at different granularities and brings mutual benefits by learning a common and rich pixel-level understanding. After pretraining on a mixed set of a limited amount of segmentation data and millions of image-text pairs, X-Decoder exhibits strong transferability to a wide range of downstream tasks in both zero-shot and finetuning settings. Notably, it achieves (1) state-of-the-art results on open-vocabulary segmentation and referring segmentation on seven datasets; (2) better or competitive finetuned performance to other generalist and specialist models on segmentation and VL tasks; and (3) flexibility for efficient finetuning and novel task composition (e.g., referring captioning and image editing shown in Fig. 1). Code, demo, video and visualization are available at: https://x-decoder-vl.github.io.

1. Introduction

Visual understanding at different levels of granularity has been a longstanding problem in the vision community. The tasks span from image-level tasks (*e.g.*, image classification [14], image-text retrieval, image captioning [8], and visual question answering (VQA) [2]), region-level localization tasks (*e.g.*, object detection and phrase grounding [58]), to pixel-level grouping tasks (*e.g.*, image in-

The work is initiated during an internship at Microsoft.

stance/semantic/panoptic segmentation [27, 35, 48]). Until recently, most of these tasks have been separately tackled with specialized model designs, preventing the synergy of tasks across different granularities from being exploited. In light of the versatility of transformers [67], we are now witnessing a growing interest in building general-purpose models that can learn from and be applied to a diverse set of vision and vision-language tasks, through multi-task learning [26, 30], sequential decoding [7, 50, 71, 80], or unified learning strategy [79, 85, 88, 89]. While these works have shown encouraging cross-task generalization capabilities, most target the unification of image-level and region-level tasks, leaving the important pixel-level understanding underexplored. In [7, 50], the authors attempt to unify segmentation into a decoding of a coordinate sequence or a color map, which, however, produces suboptimal performance and limited support for open-world generalization.

Arguably, understanding images down to the pixel level is one of the most important yet challenging problems in that: (1) pixel-level annotations are costly and undoubtedly much more scarce compared to other types of annotations; (2) grouping every pixel and recognizing them in an open-vocabulary manner is less studied; and (3) more importantly, it is non-trivial to learn from data at two substantially different granularities while also obtaining mutual benefits. Some recent efforts have attempted to bridge this gap from different aspects. In [12], Chen *et al.* propose a unified architecture Mask2Former that tackles all three types of segmentation tasks but in a closed set. To support open vocabulary recognition, a number of works study how to transfer or distill rich semantic knowledge from image-level vision-language foundation models such as CLIP [59] and ALIGN [32] to specialist models [17, 24, 60]. However, all these initial explorations focus on specific segmentation tasks of interest and do not show generalization to tasks at different granularities. In this work, we take one step further to build a generalized decoder called X-Decoder[1] towards the unification of pixel-level and image-level vision-language understanding, as shown in Figure 1.

A generalized decoding framework. We formulate all tasks including pixel-level image segmentation, image-level retrieval and vision-language tasks into a generic decoding procedure. Specifically, X-Decoder is built on top of a vision backbone and a transformer encoder for extracting multi-scale image features, following the framework of Mask2Former [12]. The key novelty lies in the decoder design. First, it takes two sets of queries as input: (i) generic non-semantic queries that aim to decode segmentation masks for universal segmentation, similar to Mask2Former [12], and (ii) newly introduced textual queries to make the decoder language-aware for a diverse set of language-related vision tasks. Second, it predicts two types of outputs: pixel-level masks and token-level semantics, and their different combinations can seamlessly support all tasks of interest. Third, we use a single text encoder to encode the textual corpus involved in all tasks, including concepts in segmentation, phrases in referring segmentation, tokens in image captioning and questions in VQA, *etc*. As a result, our X-Decoder can naturally facilitate the synergy across tasks and advocate the learning of a shared visual-semantic space, while respecting the heterogeneous nature of different tasks.

An end-to-end learning paradigm. With our generalized decoder design, we propose an end-to-end pretraining method to learn from all granularities of supervision. We unite three types of data: panoptic segmentation, referring segmentation, and image-text pairs. Unlike previous works that use pseudo-labeling techniques to extract fine-grained supervision from image-text pairs [24, 89], X-Decoder directly groups and proposes a few meaningful segmentation candidates, so that it can map the regions easily to the contents described in the captions on the fly. Meanwhile, the referring segmentation task bridges generic segmentation and image captioning by sharing the latent queries and semantic queries during decoding..

Strong zero-shot and transfer ability to a wide range of segmentation and VL tasks. Pre-trained with a limited amount of segmentation data and millions of image-text pairs (4m images), our X-Decoder supports a diversity of tasks in a zero-shot and open-vocabulary manner. Concretely, our model can be directly applied for all three types of segmentation tasks in a wide range of domains, establishing new state-of-the-art on ten settings of seven datasets. When transferred to specific tasks, our model also exhibits consistent superiority to previous works. Finally, we observe some intriguing properties in our model that it can support some novel task compositions and efficient finetuning, thanks to the flexibility endowed by our model design.

2. From Specialist to Generalist Models

2.1. Pixel-Level Understanding

Pixel-level image understanding, also known as image segmentation, has been a long-standing problem [22, 57].

Generic Segmentation. There are mainly three well-defined tasks for pixel-level understanding, including semantic [48], instance [27], and panoptic [35] segmentation. Semantic segmentation cares about the per-pixel semantic within an image [6, 11, 48], whereas instance segmentation groups pixels of the same semantic meaning into objects. Models for both tasks have evolved from CNN-based architectures [48] to transformer-based ones [11], and from two-stage models [28], one-stage models [3, 66] to the recent query-based approaches [18, 92]. With the capability of per-pixel and instance-level understanding, a natural step was taken to formulate panoptic segmentation [12, 35, 68]. Most

[1]Here, 'X' denotes versatile, and also represents 'piXel'.

recently, Mask2Former [12] proposed to address all three tasks with a unified encoder-decoder architecture. Nevertheless, all these works cope with a limited number of categories. In MSeg [38], the authors manually merge different datasets, which is still limited to being a closed set.

Open-Vocabulary Segmentation. Recently, a number of works opt to transfer or distill the rich visual-semantic knowledge from foundation models [32, 59] to specific segmentation tasks. Prominent examples include LSeg [39], OpenSeg [24], and [31]. Instead of using existing models, GroupViT [77] performed language-image pretraining from scratch with a bottom-up grouping ViT [19], while DenseCLIP [60] demonstrated the superiority of foundation models in finetuning settings compared with supervised models.

Referring Segmentation by nature is open-vocabulary. Models are usually designed specifically to learn from target datasets using various multimodal fusion strategies [29, 47, 53, 75, 83, 86]. Since the emergence of vision transformers, works like LAVT [81] enhance the cross-modal interactions from the very beginning, which led to SoTA on RefCOCO [86], RefCOCO+ [86] and G-Ref [52, 55]. CLIPSeg [51] extended the textual query to a visual query and showed superior performance not only on referring segmentation but also on semantic segmentation.

In this work, we propose X-Decoder, which is the first model to tackle generic and referring segmentation tasks all in one model. Furthermore, the generalized decoder jointly learns from segmentation data and image-text pairs end-to-end, and thus can augment the synergy across tasks for rich pixel-level and image-level understanding.

2.2. Vision-Language Understanding

Vision-language (VL) pretraining has proven to be effective for various VL tasks [41, 49, 64, 65]. The field has evolved from a transformer fusion model [10, 43, 90] with pre-extracted object features [1] to end-to-end transformers [21, 34, 40], that directly learn from raw image pixels. Recently, researchers [63,73,74] have found that image-text data at scale can be helpful for visual representation learning (*e.g.*, enabling zero-shot image classification [32, 59], action recognition [85,88], and image generation [44]). VL pre-trained models can be further extended to region-level tasks, such as phrase grounding and open-vocabulary object detection [25, 33, 54, 91], and unified frameworks that aim to combine image-text pairs with region-level data have also been proposed [4, 20, 42, 82, 89]. A comprehensive review on this topic is provided in [23].

We are clearly witnessing a trend from building specialist models to generalist ones. Early efforts [26, 30] build a multi-task learning paradigm to accommodate a diversity of tasks. However, the interactions among different tasks in these works are less studied, and the combination usually leads to performance degradation compared with specialist models. Recently, a number of works aim to reformulate the tasks into a unified sequential decoding process [7, 36, 50, 71, 80]. In this work, instead of developing a unified interface for vision and VL tasks, our X-Decoder builds a generalized decoding paradigm that can seamlessly connect the tasks by taking the common (*e.g.*, semantic) but respecting the natural differences (*e.g.*, spatial mask *v.s.* sequential language), leading to significant improvements for different segmentation and VL tasks across the board.

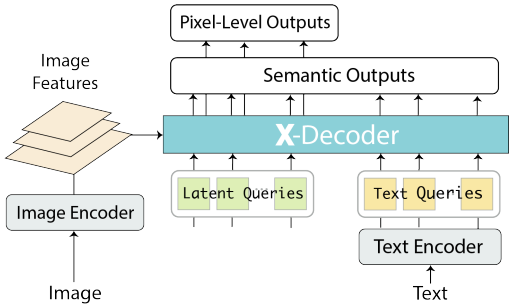

Figure 2. Overall pipeline for our model. It consists of an image encoder, a text encoder and our own designed X-Decoder.

3. X-Decoder

3.1. Formulation

Our model follows the generic design of encoder-decoder architecture as shown in Fig. 2. Given an input image $\mathbf{I} \in \mathcal{R}^{H \times W \times 3}$, we first use an image encoder $\mathbf{Enc}_I$ to extract features $\mathbf{Z}$. Afterwards, we use the text encoder $\mathbf{Enc}_T$ to encode a textual query $\mathbf{T}$ into $\mathbf{Q}_t = \langle q_t^1, \cdots, q_t^n \rangle$ of length n. The visual features, textual queries and the m non-semantic or latent queries $\mathbf{Q}_h = \langle q_h^1, \cdots, q_h^m \rangle$ are fed to our X-Decoder to predict the outputs:

$$\langle \mathbf{O}^p_{\{h,t\}}, \mathbf{O}^s_{\{h,t\}} \rangle = \mathbf{XDec}\left(\langle \mathbf{Q}_h, \mathbf{Q}_t \rangle; \mathbf{Z}\right) \tag{1}$$

where $\mathbf{O}^p_{\{h,t\}}$ and $\mathbf{O}^s_{\{h,t\}}$ are the pixel-level masks and token-level semantics for latent and textual queries, respectively. In the above formula, we note three critical designs to empower the generalization ability of our X-Decoder to a variety of vision and vision-language tasks.

We define two types of queries and outputs for X-Decoder. As discussed earlier, the queries for the decoder are categorized into latent queries $\mathbf{Q}_h$ and text queries $\mathbf{Q}_t$, which undertake generic vision and vision-language tasks, respectively. Likewise, the output is categorized into pixel-level masks and semantic embeddings. By simply using different combinations, we can adapt our X-Decoder to various tasks with the same suite of parameters.

We employ a single text encoder $\mathbf{Enc}_T$ to encode the textual corpus from all tasks. The common text encoder is used to encode referring phrases, text descriptions, image captions in the task of referring segmentation, image-text retrieval and image captioning, respectively. Furthermore, we reformulate the mask classification in segmentation into a mask-text matching problem between $\mathbf{O^s}$ and

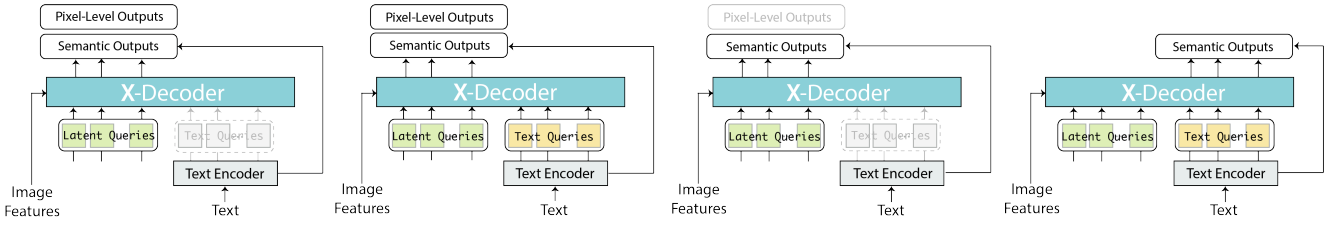

Figure 3. Unifying four different types of tasks with our proposed X-Decoder. From left to right, they are: (a) generic semantic/instance/panoptic segmentation; (b) referring segmentation; (c) image-text retrieval and (d) image captioning and VQA. The components with white text indicate not applied.

the textual embeddings of prompted textual concepts similar to [24,79]. Sharing the text encoder for all textual corpus could maximally exchange knowledge from different tasks and learn a richer and more coherent semantic space.

We fully decouple the image and text encoder. In many previous unified encoder-decoder models [7, 33, 80], the image and text are fused in the encoder side. This design makes it intractable not only for global image-text contrastive learning [59, 79], but also generative pretraining [70]. In contrast, by fully decoupling the image and text encoder and using the outputs all as queries, X-Decoder can learn from both intra-image supervisions and inter-image ones, which is essential to learn stronger pixel-level representations and support different granularity of tasks.

3.2. Unification of Tasks

Based on the above designs, X-Decoder can be used to seamlessly unify different vision and vision-language tasks, simply with different combinations of queries as inputs.

Generic Segmentation. For this task, there are no textual queries as inputs. Hence, Eq. (1) becomes:

$$\langle \mathbf{O}_h^p, \mathbf{O}_h^s \rangle = \mathbf{XDec}(\mathbf{Q}_h; \mathbf{Z}) \tag{2}$$

where $\mathbf{O}_h^p$, $\mathbf{O}_h^s$ correspond and have the same size to the latent queries $\mathbf{Q}_h$. For generic segmentation, our X-Decoder resembles Mask2former [12] but with open-vocabulary capacity since it transforms mask classification into a mask-text matching problem.

Referring Segmentation. It requires both latent and text queries as inputs:

$$\langle \mathbf{O}_h^p, \mathbf{O}_h^s \rangle = \mathbf{XDec}(\langle \mathbf{Q}_h, \mathbf{Q}_t \rangle; \mathbf{Z}) \tag{3}$$

and only uses the decoded outputs corresponding to the latent queries. Compared with Eq. (2), Eq. (3) can be considered as language-conditioned generic segmentation.

Image-Text Retrieval. The decoupled image and text encoder in our X-Decoder makes it straightforward for inter-image retrieval tasks. Specifically, we only feed the latent queries to the decoder and obtain the semantic representation of an image:

$$\mathbf{O}_h^s = \mathbf{XDec}\left(\mathbf{Q}_h; \mathbf{Z}\right) \tag{4}$$

where the last (m-th) token in $\mathbf{O}_h^s$ is used as the image representation to compute the similarities to texts.

Image Captioning and VQA. For both tasks, X-Decoder takes both latent and text queries and decodes the outputs:

$$\mathbf{O}_t^s = \mathbf{XDec}\left(\langle \mathbf{Q}_h, \mathbf{Q}_t \rangle; \mathbf{Z}\right) \tag{5}$$

where $\mathbf{O}_t^s$ correspondingly has equal size to $\mathbf{Q}_t$, and no masks are predicted. There are two slight differences between the two tasks. First, the caption prediction follows a causal masking strategy while VQA does not. Second, we use all the outputs in $\mathbf{O}_t^s$ for captioning, but only the last one to predict the answer for VQA.

The adaptation of our X-Decoder to each task is further depicted in Fig. 3. Based on this unification, we can pretrain our X-Decoder jointly with all tasks using a proper combination of queries and losses, and further finetune for individual tasks without any extra heads [2]. As discussed earlier, a lineup of works exploited a sequential decoding interface for the unification [7, 7, 13, 50, 72, 80]. However, in this work, we advocate the unification by *functionality* rather than interface, namely, we maximally share the common parts of different tasks while keeping the remaining unchanged for individual tasks.

3.3. Unified Architecture

We follow Mask2Former [12] to build our decoder architecture. Given an image $\mathbf{I} \in \mathcal{R}^{H \times W \times 3}$, we extract hierarchical visual features from L layers:

$$\mathbf{Z} = \mathbf{Enc}_l(\mathbf{I}) = \langle \mathbf{z}_l \rangle_{l=1}^{L} \tag{6}$$

where $\mathbf{z}_l \in \mathcal{R}^{H_l \times W_l \times d}$ and $\{H_l, W_l\}$ is the size of feature map at level l and d is the feature dimension. These hierarchical feature maps are important for pixel-level understanding at different scales.

One Decoder XDec for All Tasks. Given the visual features $\mathbf{Z}$, X-Decoder uses a stack of transformer layers to refine the queries and render the outputs. At layer l, it first cross-attends the visual features and then performs self-attention among latent and text queries:

$$\langle \hat{\mathbf{Q}}_h^{l-1}, \hat{\mathbf{Q}}_t^{l-1} \rangle = \mathbf{CrossAtt}(\langle \mathbf{Q}_h^{l-1}, \mathbf{Q}_t^{l-1} \rangle; \mathbf{Z}) \tag{7}$$

$$\langle \mathbf{Q}_h^{l}, \mathbf{Q}_t^{l} \rangle = \mathbf{SelfAtt}(\langle \hat{\mathbf{Q}}_h^{l-1}, \hat{\mathbf{Q}}_t^{l-1} \rangle) \tag{8}$$

In Eq. (7), we let all queries cross-attend the visual features. For latent queries, we use a masked cross-attention mechanism as in [12], and full attention for the textual queries.

[2] VQA is not used for pretraining following common practice.

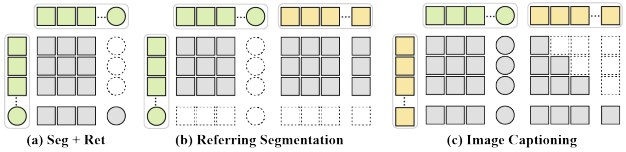

Figure 4. Interaction among latent queries (green), between latent and text queries (yellow) for (a) Generic segmentation and image/text retrieval (b) referring segmentation and (c) image captioning. The square latent query is designated for image-text retrieval.

In Eq. (8), we specifically design the self-attention mechanism: (i) we use the last latent query to extract the global image representation and the remaining for generic segmentation; (ii) for image captioning, each textual query can attend itself, its predecessors and all latent queries; (iii) for referring segmentation, latent queries attend all text queries to use it as the language condition. Based on these rules, the resulting self-attention in our X-Decoder is shown in Fig. 4.

As we illustrated in Sec. 3.2, X-Decoder always produces the masks only for the m latent queries, *i.e.*, $\mathbf{O}_h^p = \{o_1^p, \cdots, o_m^p\} \in \{0,1\}^{m\times H\times W}$ for all the latent queries. As for the semantic outputs, X-Decoder predicts the outputs for both latent and text queries, *i.e.*, $\mathbf{O}_{\{h,t\}}^s = \{o_1^s, \cdots, o_{m+n}^s\} \in \mathcal{R}^{(m+n)\times d}$, to cover both mask recognition and caption generation.

One Encoder $\mathbf{Enc}_T$ for All Texts. Given the raw text such as a phrase or caption, we convert it to discrete tokens using an off-the-shelf tokenizer and then send it to the text encoder [59]. We apply causal masking to ensure its outputs are compatible with caption decoding. For segmentation, we follow [59, 79] to convert the class name into a phrase with a text prompt (*e.g.*, "dog" $\rightarrow$ "an image of dog"), and encode the phrase as above.

3.4. End-to-End Pre-training

We train our X-Decoder in an end-to-end manner with two types of losses corresponding to the outputs.

Semantic Loss. There are three losses on the semantic outputs corresponding to three tasks. For image-text retrieval, we compute the image-language contrastive loss as [59]. We take the last valid token feature of $\mathbf{Q}_t$ from the text encoder to represent text as $\hat{q}_t$ and take the last (m-th) entry in $\mathbf{O}_h^s$ derived from X-Decoder as $\hat{o}^s$, and obtain B pairs of features for a minibatch of B image-text pairs. Afterwards, we compute the dot-product between these $B \times B$ feature pairs to obtain affinity matrix $\mathbf{S}_{it} \in \mathcal{R}^{B\times B}$, and compute the bidirectional cross-entropy loss:

$$\mathcal{L}_{it} = \mathbf{CE}(\mathbf{S}_{it}, \mathbf{y}_{it}) + \mathbf{CE}(\mathbf{S}_{it}^T, \mathbf{y}_{it}) \tag{9}$$

where $\mathbf{y}_{it}$ are the class labels corresponding to diagonal entries in $\mathbf{S}_{it}$, and $\mathbf{S}_{it}^T$ is the transpose of $\mathbf{S}_{it}$.

For mask classification, we encode all C class names including "background" into C text queries and take the last valid token feature from each to represent the concept. Afterward, we take the decoder outputs corresponding to the first $(m-1)$ latent queries and compute the dot-product between these outputs and concept embeddings to obtain an affinity matrix $\mathbf{S}_{cls} \in \mathcal{R}^{(m-1)\times C}$ and compute the loss $\mathcal{L}_{cls} = \mathbf{CE}(\mathbf{S}_{cls}, \mathbf{y}_{cls})$, with the corresponding ground-truth class $\mathbf{y}_{cls}$ guided by Hungarian Matching [5].

For image captioning, we first extract the embeddings for all tokens in the vocabulary of size V from the text encoder. Given the last n semantic outputs from X-Decoder, we compute the dot-product with all token embeddings to obtain an affinity matrix $\mathbf{S}_{cap} \in \mathcal{R}^{n\times V}$. Then we compute the cross-entropy loss $\mathcal{L}_{cap} = \mathbf{CE}(\mathbf{S}_{cap}, \mathbf{y}_{cap})$, with the ground-truth next-token id $\mathbf{y}_{cap}$.

Mask Loss. Given the $\mathbf{O}_h^p$ derived from m latent queries, we use the computed correspondence based on Hungarian Matching [5] and follow [12] to use binary cross-entropy loss $\mathcal{L}_{bce}$ and dice loss $\mathcal{L}_{dice}$ to compute the loss for masks.

Finally, we combine the above four losses to pretrain our model with segmentation and image-text pair data.

4. Experiments

Datasets and Settings. We pretrain X-Decoder on three types of data including panoptic segmentation, image-text pairs (itp), and referring segmentation. For panoptic and referring segmentation, we use COCO2017 [46] with segmentation annotations and exclude the validation sets of Ref-COCOg UMD [86] and COCO Karpathy [84]. In total, there are 104k images for segmentation pretraining, out of which 30k images are with referring segmentation annotations. For image-text pairs, we use the standard 4M corpora, including Conceptual Captions [62], SBU Captions [56], Visual Genome [37], and COCO Captions [9]. We broadly evaluate our models on all tasks covered by pretraining. In particular, we benchmark on 10 settings of 7 datasets covering a wide range of domains on zero-shot segmentation. Moreover, we finetune and report results on VQA for fine-grained visual reasoning.

Implementation Details. Our visual encoder follows [12] to use 100 latent queries and 9 decoder layers, and we add one additional latent query for image-level task. However, we do not adopt a deformable encoder as it does not generalize well to open-vocabulary settings (see in Appendix). We adopt Focal-T [78] and DaViT-B/L [16] as the vision encoder and a transformer text encoder with causal masking [59, 88] as language encoder. The models are pretrained on large-scale image-text data [88] (Base or Large) or UniCL [79] for the tiny model.

4.1. Task-Specific Transfer

Without any architecture change except adding a head for VQA, we directly finetune X-Decoder to demonstrate its task transfer capability. Table 1 presents the comparisons with previous specialized and generalized models.

Comparison with segmentation models. We list the most recent published models for individual tasks, including Mask2Former [12], Panoptic SegFormer [45], KMaX-

Method	Type	*Generic Segmentation*						*Referring*	*Retrieval*				*Captioning*		*VQA*	
		ADE			COCO			g-Ref	COCO-Karpathy		F30k-Karpathy		COCO-Karpathy		VQAv2-test	
		PQ	mAP	mIoU	PQ	mAP	mIoU	cIoU	IR@1	TR@1	IR@1	TR@1	CIDEr	BLEU	dev	std
Mask2Former (T) [12]	Segmentation	39.7	26.4	47.7	53.2	43.3	63.2	-	-	-	-	-	-	-	-	-
Mask2Former (B) [12]		⋆	⋆	53.9	56.4	46.3	67.1	-	-	-	-	-	-	-	-	-
Mask2Former (L) [12]		48.1	34.2	56.1	57.8	**48.6**	67.4	-	-	-	-	-	-	-	-	-
Pano/SegFormer (B) [45, 76]		⋆	⋆	51.0	55.4	⋆	⋆	-	-	-	-	-	-	-	-	-
kMaX-DeepLab (L) [87]		48.7	⋆	54.8	**58.1**	⋆	⋆	-	-	-	-	-	-	-	-	-
LAVT (B) [81]		-	-	-	-	-	-	61.2	-	-	-	-	-	-	-	-
UNITER (B) [10]	Vision Language (VL)	-	-	-	-	-	-	-	50.3	64.4	72.5	85.9	-	-	72.7	72.9
UNITER (L) [10]		-	-	-	-	-	-	-	52.9	65.6	75.6	87.3	-	-	73.8	74.0
VinVL (B) [90]		-	-	-	-	-	-	-	58.1	74.6	⋆	⋆	129.3	38.2	76.0	76.1
VinVL (L) [90]		-	-	-	-	-	-	-	**58.8**	75.4	⋆	⋆	130.8	38.5	76.5	76.6
ALBEF-4M (B) [40]		-	-	-	-	-	-	-	56.8	73.1	82.8	94.3	⋆	⋆	74.5	74.7
METER-Swin (B) [21]		-	-	-	-	-	-	-	54.9	73.0	79.0	92.4	⋆	⋆	76.4	76.4
UViM (L) [36]	General Purpose	⋆	⋆	⋆	45.8 [1]	⋆	⋆	-	-	-	-	-	-	-	-	-
UniT (T) [30]		-	-	-	-	-	-	-	-	-	-	-	-	-	67.6	⋆
GPV (T) [26]		-	-	-	-	-	-	-	-	-	-	-	102.3 [2]	⋆	62.5	⋆
UniTAB (B) [80]		-	-	-	-	-	-	-	-	-	-	-	119.8	36.1	70.7	71.0
Pix2Seq v2 (B) [7]		-	⋆	-	-	38.2	-	-	-	-	-	-	⋆	34.9	-	-
Unified-IO (B) [50]		-	⋆	-	-	⋆	-	-	-	-	-	-	⋆	⋆	61.8	⋆
Unified-IO (L) [50]		-	⋆	-	-	⋆	-	-	-	-	-	-	⋆	⋆	67.8	⋆
GLIPv2 (T) [89]		-	⋆	-	-	-/42.0	-	⋆	-	-	-	-	122.1	⋆	71.6	71.8
GLIPv2 (B) [89]		-	⋆	-	-	-/45.8	-	⋆	-	-	-	-	128.5	⋆	73.1	73.3
GLIPv2 (H) [89]		-	⋆	-	-	-/48.9	-	⋆	-	-	-	-	131.0	⋆	74.6	74.8
X-Decoder (T)		41.6	27.7	51.0	52.6	41.3/-	62.4	59.8 | 61.9	49.3	66.7	74.4	89.1	122.3	37.8	70.6	70.9
X-Decoder (B)		46.8	33.5	54.6	57.0	47.4/-	66.7	62.4 | 64.5	54.5	71.2	80.8	93.2	129.0	39.6	74.1	74.2
X-Decoder (L)		**49.6**	**35.8**	**58.1**	57.9	**48.6/-**	**67.8**	**64.6** | **64.6**	58.6	**76.1**	**84.4**	**94.4**	**132.1**	**40.2**	**76.8**	**77.0**

Table 1. Task-specific transfer of X-Decoder to different segmentation and VL tasks. Note: "⋆" denotes the model has the capability for the task but does not have number reported. "-" means the model does not have the ability for the specific task. "model name" means the model does not have task specific finetune. "1" is the reported pretrained number for UViM, the corresponding X-Decoder (L) has pretrained PQ 56.9. "2" is the reported coco test2014 value for GPV. "a|b" means "pretrain|finetune". "a/b" indicate "val/test".

DeepLab [87] for generic segmentation, and LAVT [81] for referring segmentation. Notably, our 25 epoch finetuned X-Decoder (L) ***establishes a new SoTA on ADE20k dataset*** that outperforms the current SoTA KMaX-DeepLab (L) on ADE Panoptic Segmentation (our model trained with 1024 resolution achieves 51.0 PQ), as well as Instance Segmentation SoTA, Mask2Former-L. On COCO, our model attains comparable or better performance to Mask2Former and kMaX-DeepLab. Finally, we compare with LAVT [81] on COCO G-ref. It is worth pointing out that with lightweight finetuning, our tiny model already outperforms LAVT-Base (**61.9** *v.s.* **61.2**). Further increasing the model size can bring additional gains by **2.6** and **2.7** points respectively, which helps to ***set a new record on this benchmark*** *in the published literature*.

Comparison with VL models. We compare with a set of VL models on image-text retrieval, image captioning and VQA in Table 1. X-Decoder achieves competitive performance across the board. Specifically, X-Decoder outperforms UNITER [10] and rivals VinVL [90] on COCO retrieval, and even beats all the baselines on Flickr30k [58]. Unlike all these works, the image and text encoders are fully decoupled in X-Decoder, which leads to a much faster inference speed. On captioning and VQA, our models also demonstrate superior performance to their counterparts. For example, it outperforms VinVL by **1.3** and **1.7** on CIDEr and BLEU, respectively. Note that most of these works use sophisticatedly designed training objectives, such as masked data modeling, image-text matching and hard-negative mining [20, 40, 69]. In contrast, X-Decoder is pretrained with image-text contrastive and image captioning, along with the segmentation losses. The simplicity and effectiveness imply a great potential for using X-Decoder as a general pretraining paradigm for VL.

Comparison with generalist models. We further compare with prior arts that explore general-purpose vision models. Limited works report the generic segmentation performance. Our model outperforms UViM [36] and Pix2Seq v2 [7] significantly on COCO panoptic (**56.9** *v.s.* **45.8**) and instance segmentation (**46.7** ***v.s.*** **38.2**), respectively (The X-Decoder (L) have the same zero-shot and finetuning performance). With the same amount of segmentation data, these margins strongly justify our model design, *i.e.*, unifying functionality *without* any tweaks for individual tasks. When compared with GLIPv2 [89], our model achieves comparable performance. Note that GLIPv2 uses over 10M pretraining data, including around 2M with box supervision. Despite the huge gap in pretraining data, X-Decoder outperforms GLIPv2 on both captioning and VQA.

Efficient Finetuning. Finally, we study whether our pretrained X-Decoder can be finetuned for segmentation with a low cost. In Table 3, we show that we can simply finetune the class embedding layer, mask embedding layer or the whole decoder to reach a decent segmentation performance and surpass the fully finetuned tiny SoTA models like kMaX-DeepLab [87]. These results imply an efficient way of using our pretrained X-Decoder models.

4.2. Zero-Shot Transfer

Without any change in model weights, X-Decoder can be directly applied to various segmentation tasks and datasets after pretraining. In Table 2, we evaluate our model in a zero-shot manner on seven commonly used segmentation

Model	COCO (p/s)			ITP	Fix	EM	Pse-	ADE-150			A-847	VOC	PC-59	PC-459	SUN	SCAN-20		SCAN-41	Cityscapes			BDD	
	m	cls	cap				udo	PQ	mAP	mIoU	mIoU	mIoU	mIoU	mIoU	mIoU	mIoU	PQ	mIoU	mIoU	mAP	PQ	mIoU	PQ
MSeg (B) [38]	✓	✓	✗	✗	✗	✗	✗	33.7	32.6	19.1	⋆	73.4	43.4	⋆	29.6	33.4	⋆	⋆	46.9	24.8	51.1	44.9	⋆
GroupViT (S)	✗	✗	✗	✓	✗	✓	✗	-	-	⋆	⋆	52.3	22.4	⋆	⋆	⋆	-	⋆	⋆	-	-	⋆	-
LSeg+ (B) [39]	✓	✓	✗	✗	✓	✓	✗	-	-	18.0	3.8	⋆	46.5	7.8	⋆	⋆	-	⋆	⋆	-	-	⋆	-
ZegFormer (B) [15]	✓	✓	✗	✗	✓	✓	✗	-	-	⋆	8.1	80.7	⋆	⋆	⋆	⋆	-	⋆	⋆	-	-	⋆	-
OpenSeg (B) [25]	✓	✗	✓	✗	✓	✓	✓	-	-	21.1	6.3	70.3	45.9	9.0	⋆	⋆	-	⋆	⋆	-	-	⋆	-
OpenSeg (B) [25]	✓	✗	✓	✓	✓	✓	✓	-	-	26.4	8.1	70.2	44.8	11.5	⋆	⋆	-	⋆	⋆	-	-	⋆	-
MaskCLIP (L) [17]	✓	✓	✗	✗	✓	✓	✗	15.1	6.0	23.7	8.2	⋆	45.9	10.0	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆
X-Decoder-Seg (B)	✓	✓	✗	✗	✗	✗	✗	15.3	8.3	19.5	2.9	95.7	63.5	13.3	33.0	41.6	32.5	22.4	47.3	22.8	35.2	44.1	14.1
X-Decoder-Seg$^+$ (B)	✓	✓	✓	✗	✗	✗	✗	16.9	9.5	23.8	4.6	97.8	64.7	12.1	32.2	35.1	33.8	18.5	47.6	**25.9**	36.9	42.7	16.6
X-Decoder (T)	✓	✓	✓	✓	✗	✗	✗	18.8	9.8	25.0	6.4	96.2	62.9	12.3	34.5	37.8	30.7	21.7	47.3	16.0	37.2	42.4	16.4
X-Decoder (B)	✓	✓	✓	✓	✗	✗	✗	21.1	11.7	27.2	8.2	**97.9**	**65.1**	14.7	39.6	40.3	35.4	24.8	50.8	22.3	**39.5**	45.1	17.1
X-Decoder (L)	✓	✓	✓	✓	✗	✗	✗	**21.8**	**13.1**	**29.6**	**9.2**	97.7	64.0	**16.1**	**43.0**	**49.5**	**39.5**	**29.7**	**52.0**	24.9	38.1	**47.2**	**17.8**

Table 2. One suite of model weights for open-vocabulary image segmentation. Note: "ITP" means image-text pairs. "Fix" indicates whether contains fixed text/image encoder. "EM" means whether the model has extra modules that are designed for open-vocabulary settings (e.g. Adaptor, class agnostic proposal, and etc.). "Pseudo" means whether the method uses an extra step to extract pseudo label image-text pairs. "gray" color means a fully supervised approach. "light purple" color means a semi-supervised learning approach. "FL-in21k" means the backbone is pretained with in21k data using a FocalNet backbone. For COCO, different methods use different supervisions of mask (m), class label (cls) and caption (cap). "⋆ and -" follows Table 1

Method	C.E.	M.E	Q.E	Dec.	#Param	ADE			Cityscapes		
						PQ	mAP	mIoU	PQ	mAP	mIoU
Mask2Former (T) [12]	-	-	-	-	-	39.7	26.4	47.7	63.9	39.1	80.5
Pano/SegFormer (T) [45, 76]	-	-	-	-	-	36.4	⋆	46.5	⋆	⋆	⋆
kMaX-DeepLab (T) [87]	-	-	-	-	-	41.5	⋆	45.0	64.3	38.5	79.7
Mask2Former (S) [12]	-	-	-	-	-	⋆	⋆	51.3	64.8	40.7	81.8
Mask2Former (B) [12]	-	-	-	-	-	⋆	⋆	53.9	66.1	42.8	82.7
Mask2Former (L) [12]	-	-	-	-	-	48.1	34.9	56.1	66.6	43.6	82.9
X-Decoder (L)	✓	✗	✗	✗	0.26M	44.3	33.2	54.6	65.1	41.4	81.7
	✓	✓	✗	✗	1.05M	43.9	33.2	53.9	64.8	41.2	81.2
	✓	✓	✓	✗	1.15M	44.0	32.8	54.0	64.6	41.1	81.5
	✓	✓	✓	✓	38.3M	**47.0**	**35.1**	**56.0**	**65.6**	**42.2**	**81.7**

Table 3. Performance with different efficient finetuning strategies for X-Decoder large, and comparisons with fully-finetuned models. Note: C.M denotes class embedding, M.E. denotes mask embedding, Q.E. denotes query embedding, #Param means the number of parameters tuned.

datasets in 10 different settings from diverse domains, including common indoor, outdoor and self-driving scenarios. We report PQ, mAP and mIoU for generic segmentation quantitatively, and qualitatively show examples on various dataset in the Appendix.

Comparison with baselines. We build two X-Decoder variants: (1) X-Decoder-Seg, which is only trained with COCO panoptic segmentation using a text encoder for class names; and (2) X-Decoder-Seg$^+$, where we take the heuristic way to extract noun phrases from COCO captions and use them as extra supervision on top of the matched decoder outputs. First, X-Decoder-Seg shows clear advantages on open-vocabulary segmentation over MSeg [38], that manually conducts label mapping across different datasets. Second, the extra supervision from COCO captions improves model performance on 9 out of 15 metrics, which indicates the benefit of joint learning with image-level supervision. Third, when pretraining with the full X-Decoder, the performance is significantly boosted. Notably, the mIoU metric is improved by **7.4**, **3.4** and **2.6** on SUN, ADE-150 and PC-459, respectively.

Comparison with state-of-the-art. We further compare with the most advanced methods for open-vocabulary image segmentation in Table 2. Clearly, our models achieve the best results across all datasets. Among the base-sized models, X-Decoder (B) outperforms OpenSeg (B) [24] on two challenging datasets, ADE-150 and PC-459 for semantic segmentation. Scaling X-Decoder to large size further improves mIoU by **2.4** and **1.4** on these two datasets. Among prior arts, MaskCLIP [17] is the first proposed for open-vocabulary panoptic segmentation by combining Mask2Former with CLIP models. With COCO caption supervisions, our simple baseline X-Decoder-Seg$^+$ already performs comparably. The full version of our tiny model X-Decoder (T) surpasses MaskCLIP across the board except A-847. We note that these comparisons are not strictly fair in terms of supervision, settings and models used. However, these results demonstrate the effectiveness of our X-Decoder to learn from the different granularity of supervisions *end-to-end* for open-vocabulary segmentation, which leads to ***new SoTA on 10 settings of 7 datasets across three segmentation tasks***.

4.3. Model Inspection

Pretraining Tasks. By default, we exploit four pre-training tasks including generic and referring segmentation, captioning and retrieval. In Table 6, we keep the generic segmentation while ablating the importance of the other pre-training tasks. Accordingly, we have the following observations:

Image-text retrieval help open-vocabulary segmentation: On ADE, mIoU decreases from 23.4 to 21.8 and PQ by 0.7 without image-text retrieval. As both tasks share semantic space, improved visual-semantic alignment boosts recognition of novel concepts.

Image captioning helps referring segmentation and vice versa: COCO g-Ref drops **2.0** pts without training with image captioning, and CIDEr falls **3.2** pts without training with referring tasks. This indicate sharing a text encoder and joint training enhances text input understanding.

Image captioning and retrieval can mutually benefit each other: Removing captioning in pretraining, R@1 drops by **0.8**; and CIDEr decreases **3.2** pts without retrieval task. It indicates X-Decoder fosters generative and contrastive learning synergy.

Query Interactions. The interaction among tasks is highly dependent on the interaction between latent and text

Model	COCO			ADE			COCO-Karparthy			g-Ref
	PQ	mAP	mIoU	PQ	mAP	mIoU	IR@1	IR@1	CIDEr	cIoU
X-Decoder	**51.4**	**40.5**	**62.8**	14.7	**9.6**	**23.4**	30.7	48.5	**82.0**	**59.7**
* text: [yny]	51.4	39.8	61.7	14.7	9.4	22.2	29.9	46.9	78.6	57.7
* text: [nyy]	51.4	38.6	61.7	15.2	9.4	23.1	30.3	47.5	78.9	59.4
* latent: [yyn]	50.9	39.6	62.0	**15.5**	9.4	22.8	29.8	47.6	81.1	57.6

Table 4. Ablation of query interaction in X-Decoder. [x,x,x] denotes whether attend [object latent query, image latent query, text query]

Model	COCO			ADE			COCO-Karparthy			g-Ref
	PQ	mAP	mIoU	PQ	mAP	mIoU	IR@1	TR@1	CIDEr	cIoU
X-Decoder	**51.4**	**40.5**	62.8	14.7	**9.6**	**23.4**	**30.7**	**48.5**	**82.0**	**59.7**
- Retrieval	51.4	40.4	62.6	*14.0*	9.2	*21.8*	n/a	n/a	*78.8*	59.2
- Captioning	*51.1*	40.4	**63.2**	15.0	9.6	23.2	*29.9*	48.1	n/a	*57.7*
- Referring	*51.1*	*39.7*	*62.3*	**15.2**	*8.9*	22.6	30.0	*47.6*	*78.8*	n/a

Table 6. Ablation of pretraining tasks by removing one at a time. We bold the best entry and underline the worst entry in each column.

Model	COCO			ADE			COCO-Karparthy			g-Ref
	PQ	mAP	mIoU	PQ	mAP	mIoU	IR@1	TR@1	CIDEr	cIoU
* bs 1024	50.9	39.5	62.4	15.2	10.0	24.6	30.6	48.1	85.0	58.0
* bs 768	51.0	39.5	62.4	15.4	10.0	24.2	29.0	46.8	78.6	58.8
* bs 512	50.7	39.3	62.0	14.9	9.7	24.3	27.4	43.8	76.1	58.6

Table 5. Ablation of VL batch size. We mark the significant drop metrics in green.

Model	COCO			ADE			COCO-Karparthy			g-Ref
	PQ	mAP	mIoU	PQ	mAP	mIoU	IR@1	TR@1	CIDEr	cIoU
Full Datasets	50.9	39.5	62.4	15.2	10.0	24.6	30.6	48.1	85.0	58.0
- coco	50.9	39.9	62.2	15.3	9.8	24.4	27.4	38.2	32.6	59.4
- cc3m	51.2	39.7	62.6	15.5	10.1	24.6	31.0	50.0	81.2	58.3
- vg	51.1	39.8	62.4	14.6	9.7	23.8	36.1	56.1	107.1	58.3
- sbu	51.1	39.8	62.4	15.3	9.5	24.6	30.3	48.3	81.2	58.3

Table 7. Ablation of VL datasets in X-Decoder. A single VL dataset is removed in each row. And we mark the metrics that significantly drop/increase in green/red.

Figure 5. An example of region retrieval as a showcase of task composition of image-text retrieval and referring segmentation.

queries. We have described how the queries interact with each other by default in Fig. 4. Here, we investigate how our model behaves with different interactions in Tab. 4:

Image captioning requires both fine-grained and global image information: Comparing rows 1-3 in Tab. 4, CIDEr score drops significantly when information flow from global latent queries or other latent queries to text queries is cut off (**82.0** $\rightarrow$ **78.6** and **78.9**, respectively).

Language-condition is important for ref-segmentation: In the last row of Tab. 4, turning off text-to-latent query interaction significantly decrease ref-segmentation (**59.7** $\rightarrow$ **57.6**) performance, indicating generic segmentation can't be converted to referring segmentation using post-hoc matching with referring texts easily.

VL Batch Size & Dataset The default batch size of VL task is 1024, we explore the gradual decreasing of VL batch size in Tab. 5. In addition, each VL dataset is removed to investigate the pre-trained performance on different tasks.

Decreasing VL batch size hurts VL tasks and open-vocab Segmentation performance: As shown in Tab. 5, decreasing the VL task batch size from 1024 to 256 significantly hurts the retrieval and captioning performance as well as minor influence on open-vocabulary settings.

VG dataset hurts pretraining VL tasks performance but improves open-vocab segmentation: As shown in Table 7, removing the visual genome from the pretraining VL dataset significantly improves captioning task with 22.1 points in pretraining caused by the different annotation style of that dataset.

4.4. Task Composition

X-Decoder has the unique benefit of task interaction. We demonstrate our model can perform region-based retrieval (Fig. 5) and referring based captioning (Fig. 1) without any architecture/weight change. As shown in Fig. 5, given a set of animal images and text query, our model first retrieves the correct image and then grounds the query in pixel level. Further, our model can be easily adapted to referring captioning by localizing a given word and then modulating the predicted mask in the cross-attention layers for text queries. This will allow the text queries to focus on the grounded region only. Thus lead to the generated caption that focus on the specific area. Lastly, we also integrate X-Deocder with diffusion model for referring image editing and inpainting. This has been demonstrated in Fig. 1.

5. Conclusion

We introduce X-Decoder, a versatile model for pixel and image-level vision-language understanding. Its unified design enables generic segmentation, referring segmentation, and VL tasks with strong generalizability and SoTA/Comparable performance. We hope this work can shed a light on the design of the next-generation general-purpose vision system.

Acknowledgements. We thank Haotian Zhang for constructive discussions. This work is supported in part by NSF IIS2150012, WARF, and the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00871, Development of AI Autonomy and Knowledge Enhancement for AI Agent Collaboration).

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015.

[3] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9157–9166, 2019.

[4] Zhaowei Cai, Gukyeong Kwon, Avinash Ravichandran, Erhan Bas, Zhuowen Tu, Rahul Bhotika, and Stefano Soatto. X-detr: A versatile architecture for instance-wise vision-language tasks. *arXiv preprint arXiv:2204.05626*, 2022.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[7] Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J Fleet, and Geoffrey Hinton. A unified sequence interface for vision tasks. *arXiv preprint arXiv:2206.07669*, 2022.

[8] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[9] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[10] 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.

[11] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022.

[12] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022.

[13] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[15] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11583–11592, 2022.

[16] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. *arXiv preprint arXiv:2204.03645*, 2022.

[17] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary panoptic segmentation with maskclip. *arXiv preprint arXiv:2208.08984*, 2022.

[18] Bin Dong, Fangao Zeng, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Solq: Segmenting objects by learning queries. *Advances in Neural Information Processing Systems*, 34:21898–21909, 2021.

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[20] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, Jianfeng Gao, and Lijuan Wang. Coarse-to-fine vision-language pre-training with fusion in the backbone. In *NeurIPS*, 2022.

[21] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of training end-to-end vision-and-language transformers. In *CVPR*, 2022.

[22] King-Sun Fu and JK Mui. A survey on image segmentation. *Pattern recognition*, 13(1):3–16, 1981.

[23] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. Vision-language pre-training: Basics, recent advances, and future trends. *arXiv preprint arXiv:2210.09263*, 2022.

[24] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Open-vocabulary image segmentation. *arXiv preprint arXiv:2112.12143*, 2021.

[25] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

[26] Tanmay Gupta, Amita Kamath, Aniruddha Kembhavi, and Derek Hoiem. Towards general purpose vision systems: An end-to-end task-agnostic vision-language architecture. In *CVPR*, 2022.

[27] Abdul Mueed Hafiz and Ghulam Mohiuddin Bhat. A survey on instance segmentation: state of the art. *International journal of multimedia information retrieval*, 9(3):171–189, 2020.

[28] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[29] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *European Conference on Computer Vision*, pages 108–124. Springer, 2016.

[30] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1439–1449, 2021.

[31] Dat Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, and Ehsan Elhamifar. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7031, 2022.

[32] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021.

[33] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021.

[34] Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021.

[35] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019.

[36] Alexander Kolesnikov, André Susano Pinto, Lucas Beyer, Xiaohua Zhai, Jeremiah Harmsen, and Neil Houlsby. UViM: A unified modeling approach for vision with learned guiding codes. *arXiv preprint arXiv:2205.10337*, 2022.

[37] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[38] John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. Mseg: A composite dataset for multi-domain semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2879–2888, 2020.

[39] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022.

[40] Junnan Li, Ramprasaath R Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021.

[41] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint*, 2019.

[42] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.

[43] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020.

[44] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. *arXiv preprint arXiv:2301.07093*, 2023.

[45] Zhiqi Li, Wenhai Wang, Enze Xie, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, Ping Luo, and Tong Lu. Panoptic segformer: Delving deeper into panoptic segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1280–1289, 2022.

[46] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[47] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1271–1280, 2017.

[48] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[49] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019.

[50] 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.

[51] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7086–7096, 2022.

[52] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

[53] Edgar Margffoy-Tuay, Juan C Pérez, Emilio Botero, and Pablo Arbeláez. Dynamic multimodal instance segmentation guided by natural language queries. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 630–645, 2018.

[54] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection with vision transformers. *arXiv preprint arXiv:2205.06230*, 2022.

[55] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision*, pages 792–807. Springer, 2016.

[56] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned pho-

tographs. *Advances in neural information processing systems*, 24, 2011.

[57] Nikhil R Pal and Sankar K Pal. A review on image segmentation techniques. *Pattern recognition*, 26(9):1277–1294, 1993.

[58] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[59] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[60] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022.

[61] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[62] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[63] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *CVPR*, 2022.

[64] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: Pre-training of generic visual-linguistic representations. In *ICLR*, 2019.

[65] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019.

[66] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 282–298. Springer, 2020.

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[68] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5463–5474, 2021.

[69] Jianfeng Wang, Xiaowei Hu, Zhe Gan, Zhengyuan Yang, Xiyang Dai, Zicheng Liu, Yumao Lu, and Lijuan Wang. Ufo: A unified transformer for vision-language representation learning. *arXiv preprint arXiv:2111.10023*, 2021.

[70] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.

[71] 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.

[72] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*, 2022.

[73] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

[74] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

[75] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4984, 2022.

[76] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *arXiv preprint arXiv:2105.15203*, 2021.

[77] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022.

[78] Jianwei Yang, Chunyuan Li, and Jianfeng Gao. Focal modulation networks. *arXiv preprint arXiv:2203.11926*, 2022.

[79] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *CVPR*, 2022.

[80] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *European Conference on Computer Vision*, pages 521–539. Springer, 2022.

[81] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022.

[82] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang

Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *arXiv preprint arXiv:2209.09407*, 2022.

[83] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10502–10511, 2019.

[84] Xuwang Yin and Vicente Ordonez. Obj2text: Generating visually descriptive language from object layouts. *arXiv preprint arXiv:1707.07102*, 2017.

[85] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[86] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016.

[87] Qihang Yu, Huiyu Wang, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. k-means mask transformer. In *European Conference on Computer Vision*, pages 288–307. Springer, 2022.

[88] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[89] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *arXiv preprint arXiv:2206.05836*, 2022.

[90] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. VinVL: Revisiting visual representations in vision-language models. In *CVPR*, 2021.

[91] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022.

[92] Xueyan Zou, Haotian Liu, and Yong Jae Lee. End-to-end instance edge detection. *arXiv preprint arXiv:2204.02898*, 2022.