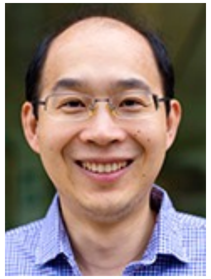
Paper tag: THU-AM-208

PolyFormer: Referring Image Segmentation as Sequential Polygon Generation




Jiang Liu[1*†] Hui Ding[2*] Zhaowei Cai[2] Yuting Zhang[2] Ravi Kumar Satzoda[2] Vijay Mahadevan[2] R. Manmatha[2]

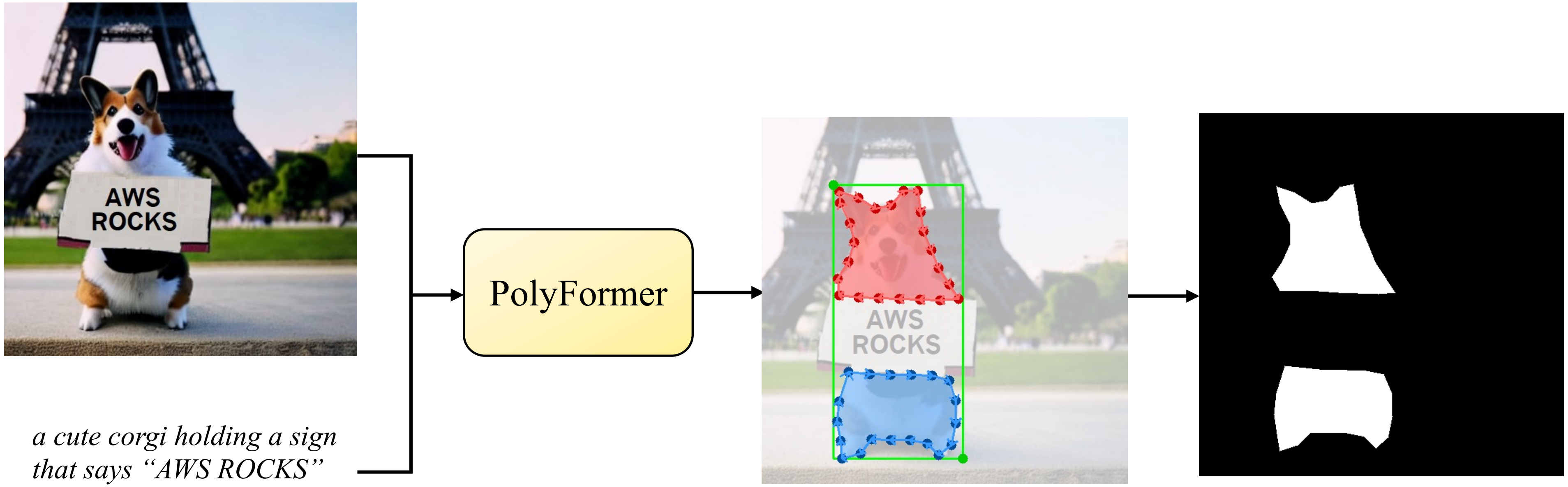
[1] Johns Hopkins University, [2] AWS AI Labs, [*]Equal Contribution, [†]Work done during internship at AWS AI Labs

https://polyformer.github.io/

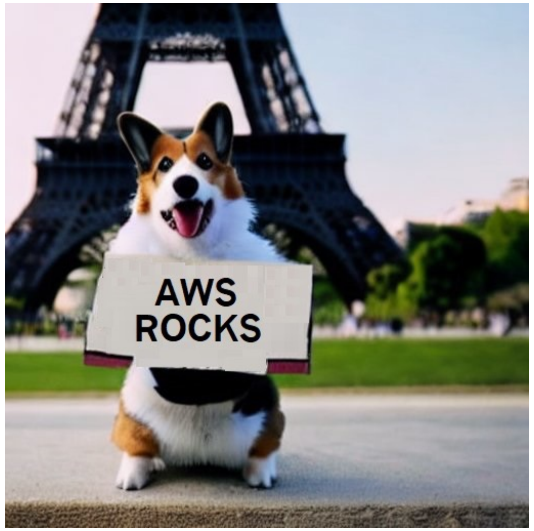
Code & Demo Available

PolyFormer Overview

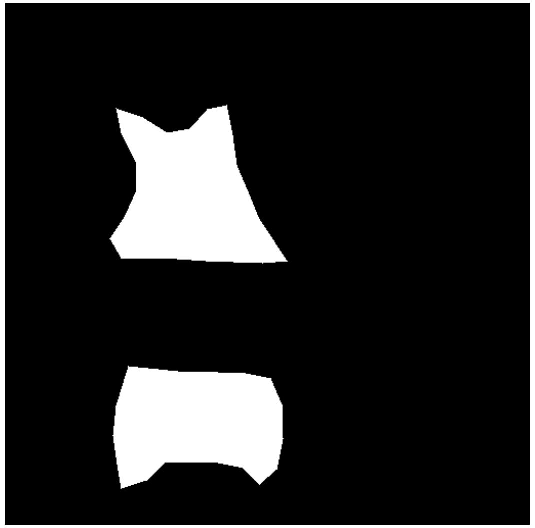
- **Unified framework** for referring image segmentation and referring expression comprehension
- **Regression-based decoder** for accurate coordinate prediction
- **Superior performance** across all main referring image segmentation



Referring Image Segmentation



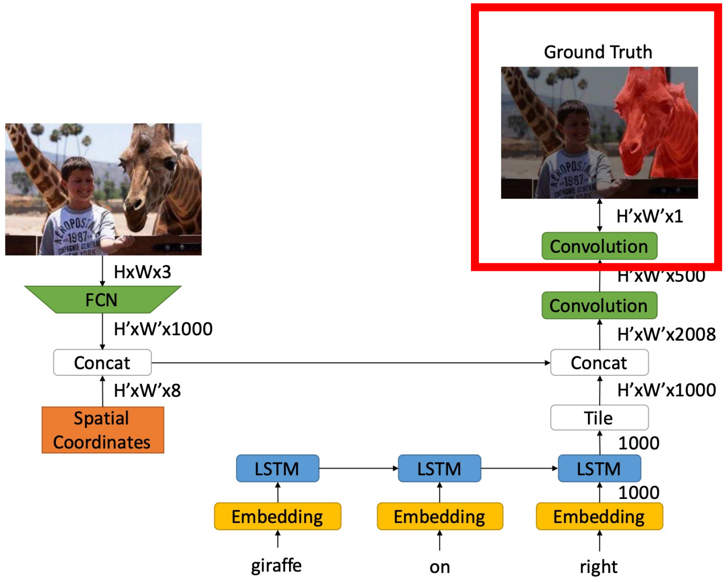
RIS



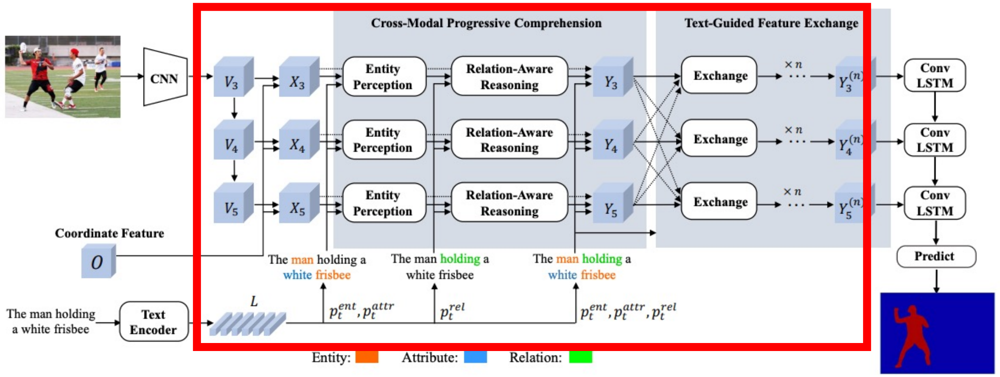
a cute corgi holding a sign
that says "AWS ROCKS"

Existing Work

- **Mask-based** dense prediction
  - Neglect the structure among the output predictions
- Complex **multi-modal feature fusion**



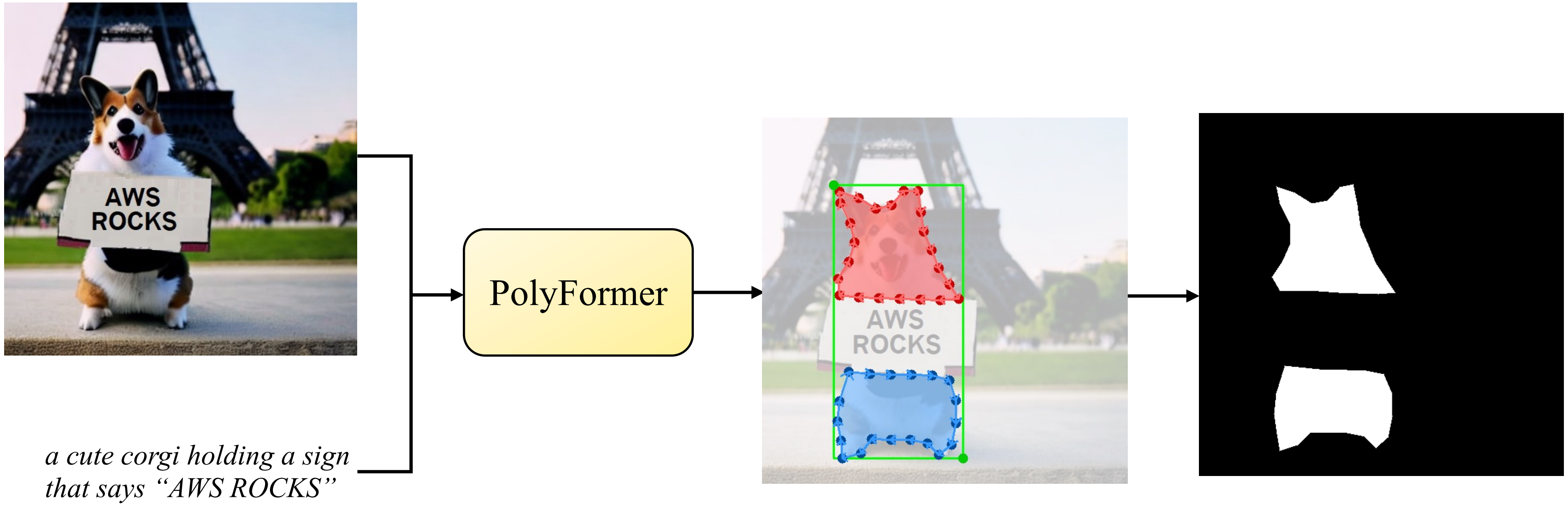
RMI [Liu et al, ICCV2017]



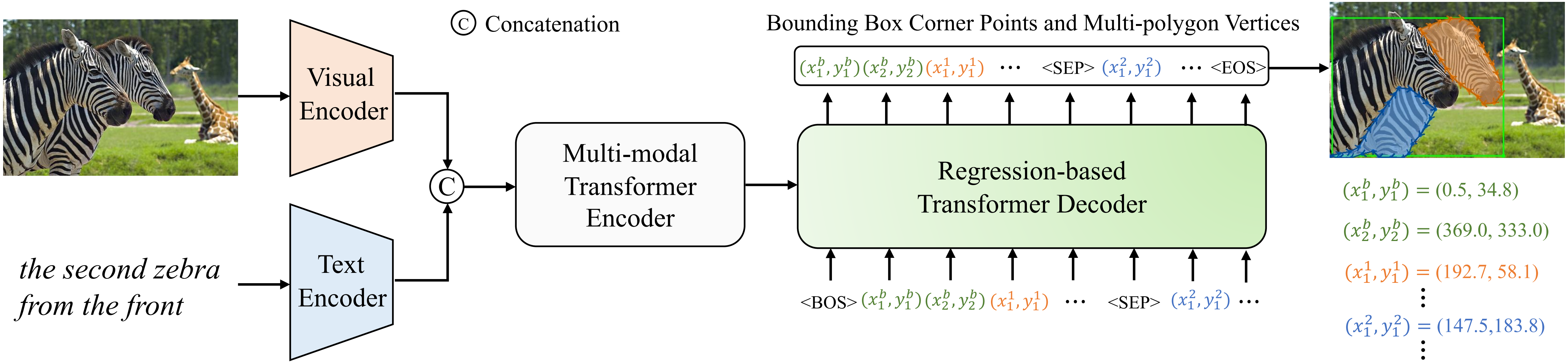
Huang et al. "Referring Image Segmentation via Cross-Modal Progressive Comprehension." CVPR2020

PolyFormer

- Sequence-to-sequence formulation

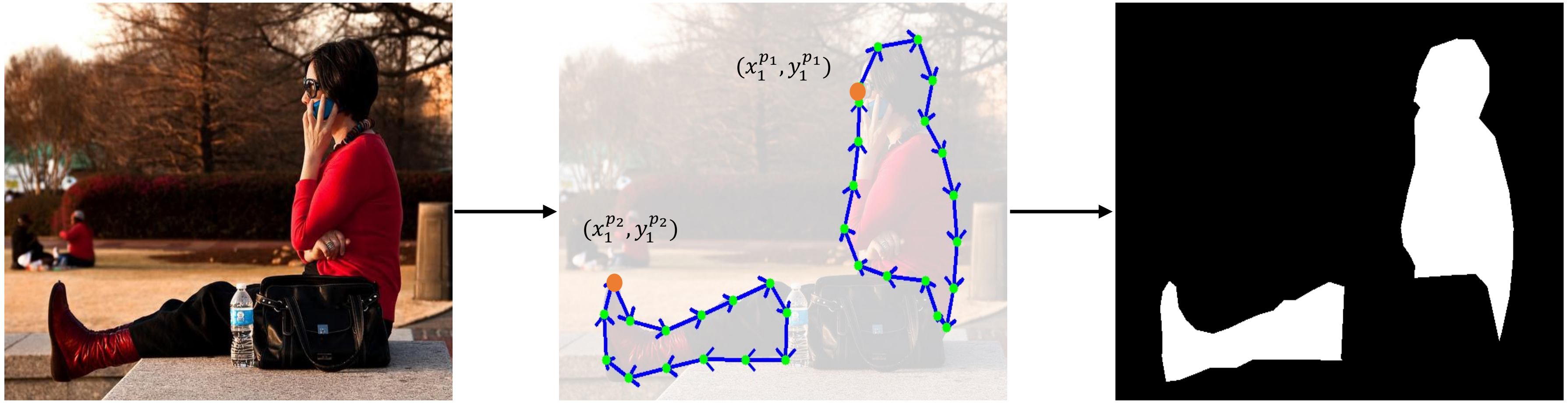


Model Architecture



Target Sequence Generation

- Polygon ordering
  - Start from top-left
  - Clockwise direction

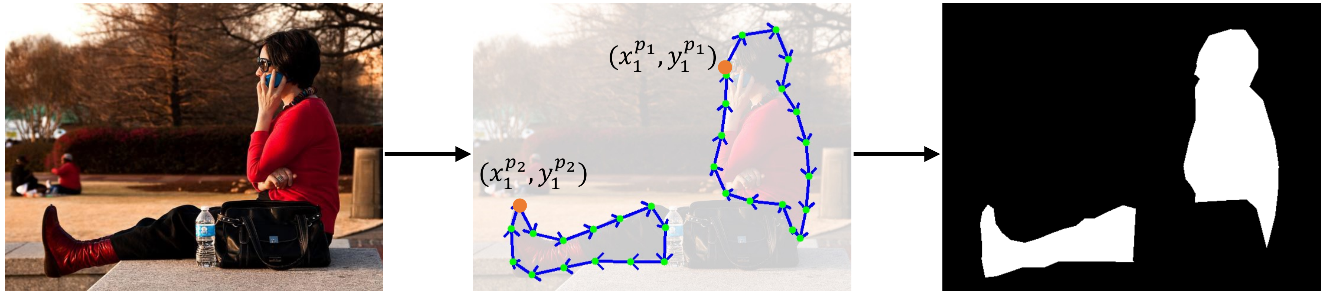


$$(x_1, y_1)\ ,(x_2, y_2)\ , \ldots\ldots\ , (x_K, y_K)$$

Target Sequence Generation

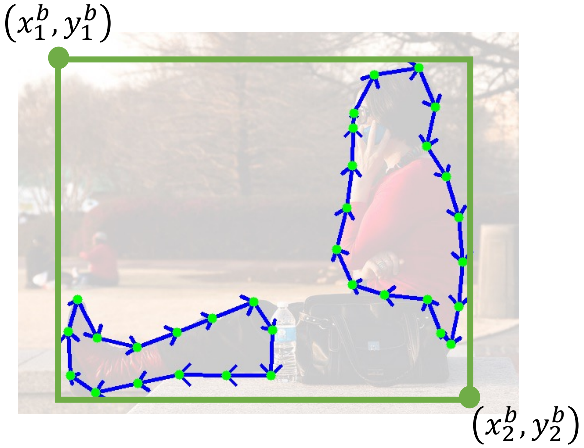
- Multi-polygon case
  - Separator token <SEP>

$$\underbrace{(x_1^{p_1}, y_1^{p_1}) \;\ldots\ldots\; (x_{n_1}^{p_1}, y_{n_1}^{p_1})}_{\text{Polygon 1}} \;\boxed{\textbf{<SEP>}}\; \underbrace{(x_1^{p_2}, y_1^{p_2}) \;\ldots\ldots\; (x_{n_2}^{p_2}, y_{n_2}^{p_2})}_{\text{Polygon 2}} \;\ldots\ldots$$



Target Sequence Generation

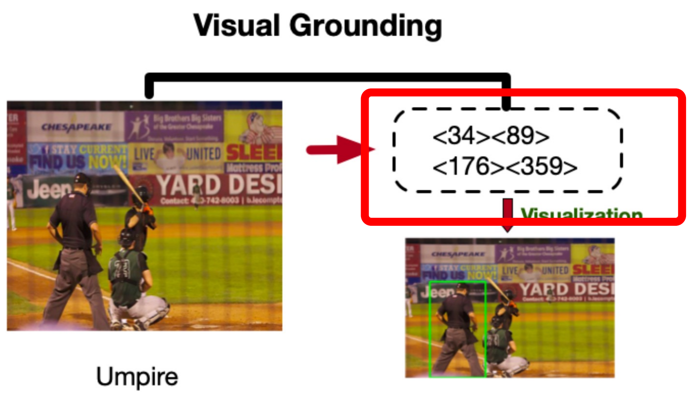
- Unified sequence with bounding box
  - Bounding box: $(x_1^b, y_1^b), (x_2^b, y_2^b)$
- Final target sequence:



<BOS> $\underbrace{(x_1^b, y_1^b), \quad (x_2^b, y_2^b)}_{\text{Bounding box}}$ $\underbrace{(x_1^{p_1}, y_1^{p_1}) \ \ldots\ldots \ (x_{n_1}^{p_1}, y_{n_1}^{p_1})}_{\text{Polygon 1}}$ <SEP> $\underbrace{(x_1^{p_2}, y_1^{p_2}) \ \ldots\ldots \ (x_{n_2}^{p_2}, y_{n_2}^{p_2})}_{\text{Polygon 2}}$ ...... <EOS>

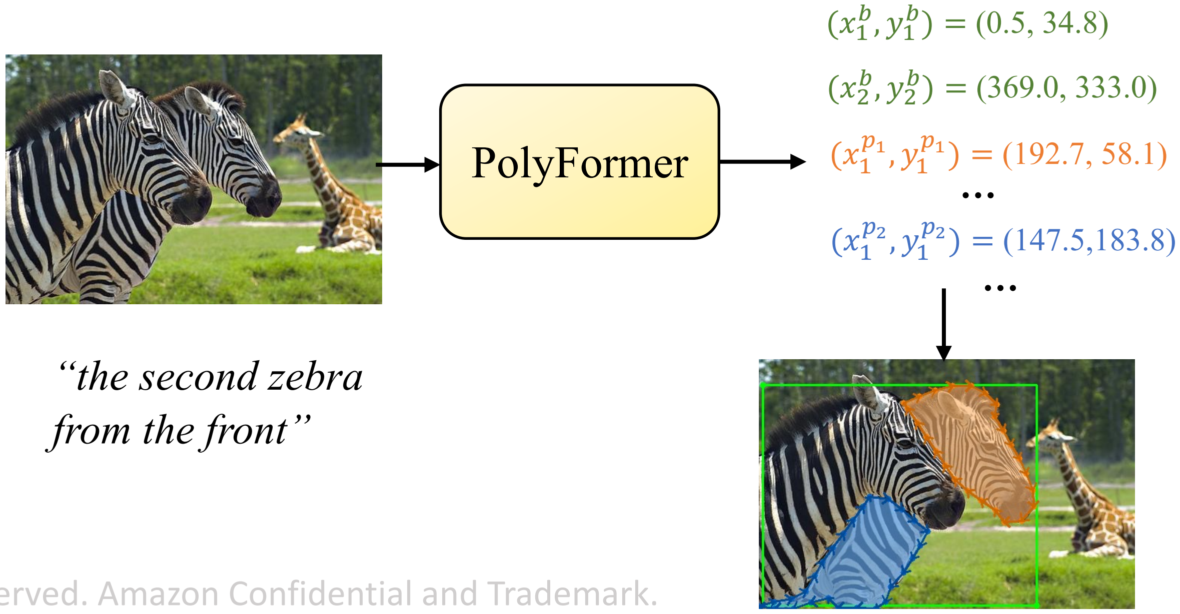
Regression-based Decoder

- Previous Seq2Seq framework: Coordinate prediction as a ***classification task***
  - *Continuous* coordinates => *discrete* bins
  - quantization error
  - Inaccurate supervision



OFA [Wang et al, ICML2022]

- PolyFormer: geometric localization as a ***regression*** task
  - Directly predict floating-point coordinate
  - No quantization error
  - Accurate localization

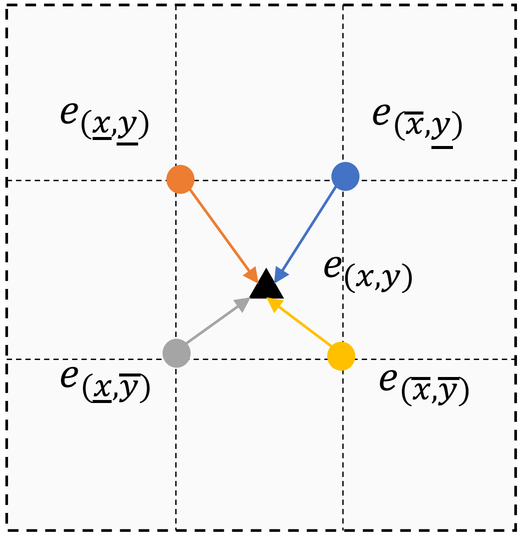


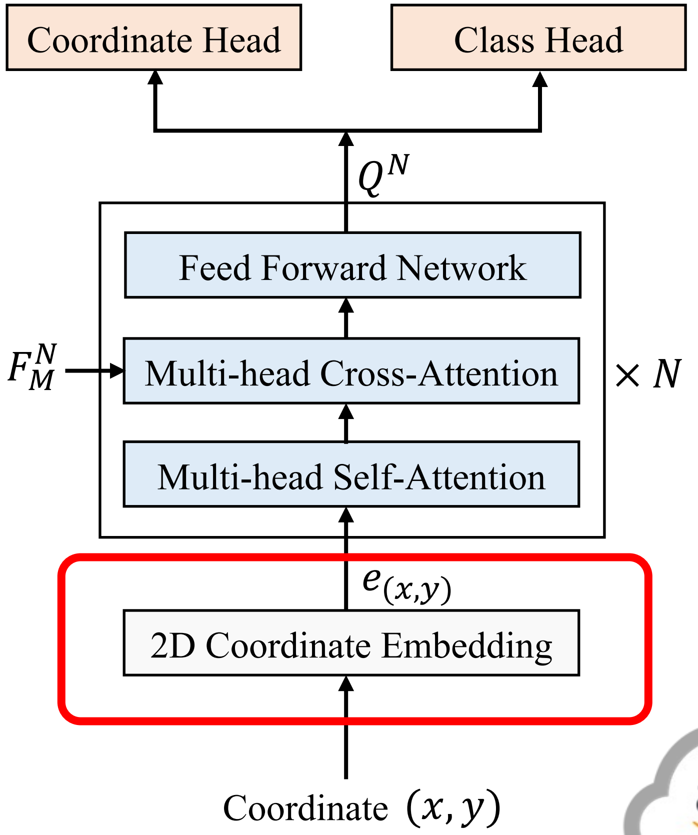
Regression-based Transformer Decoder

- 2D Coordinate Embedding

$$
\begin{aligned}
e_{(x,y)} = & (\bar{x}-x)(\bar{y}-y)\cdot e_{(\underline{x},\underline{y})} + (x-\underline{x})(\bar{y}-y)\cdot e_{(\bar{x},\underline{y})} + \\
& (\bar{x}-x)(y-\underline{y})\cdot e_{(\underline{x},\bar{y})} + (x-\underline{x})(y-\underline{y})\cdot e_{(\bar{x},\bar{y})}.
\end{aligned}
$$

2D Coordinate Embedding Codebook $\mathcal{D} \in \mathbb{R}^{B_H \times B_W \times C_e}$





Regression-based Transformer Decoder

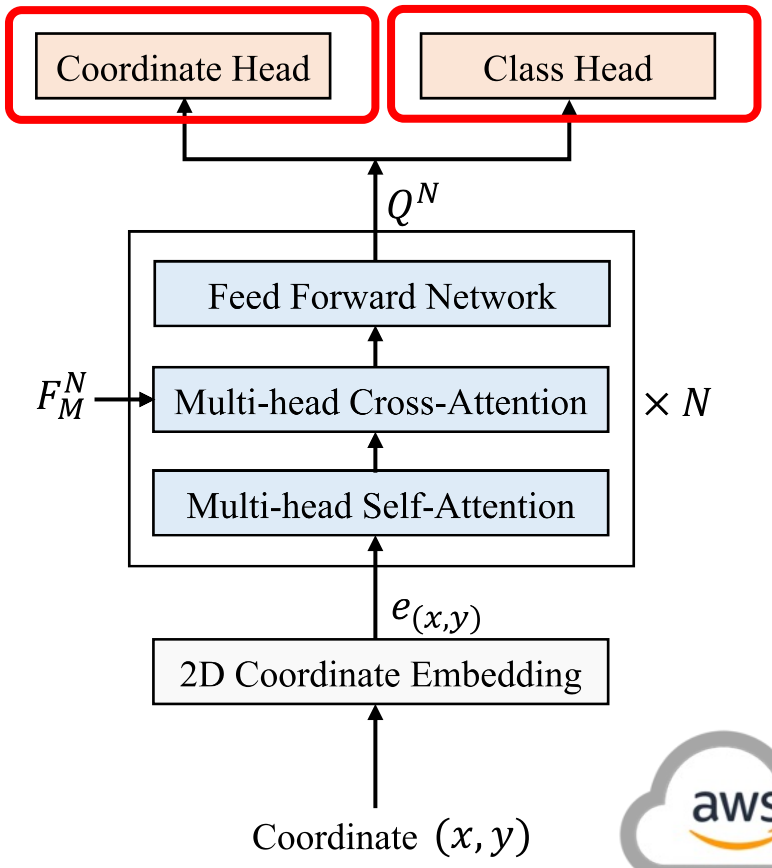
- Prediction Heads
  - **Coordinate head**
    - 3-layer feed-forward network (FFN)

$$(\hat{x}, \hat{y}) = Sigmoid(FFN(Q^N)).$$

  - **Class head**
    - Linear classification layer

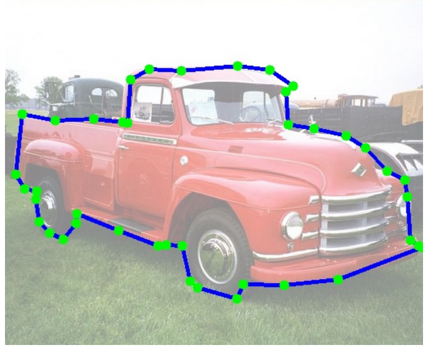
$$\hat{p} = W_c Q^N + b_c,$$

    - Separator token <SEP>, coordinate token <COO>, end-of-sequence token <EOS>

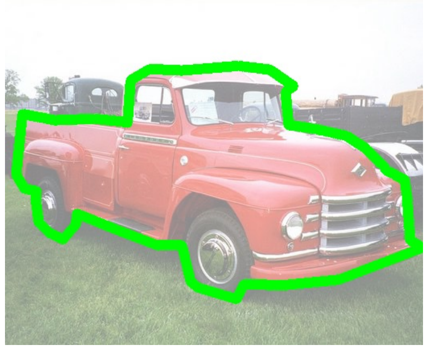


Training: Polygon Augmentation

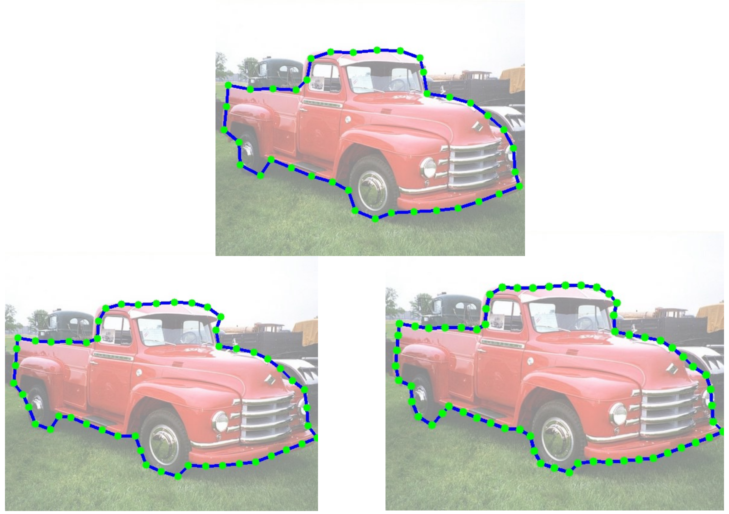
polygons at different levels of granularity



(a) Original polygon



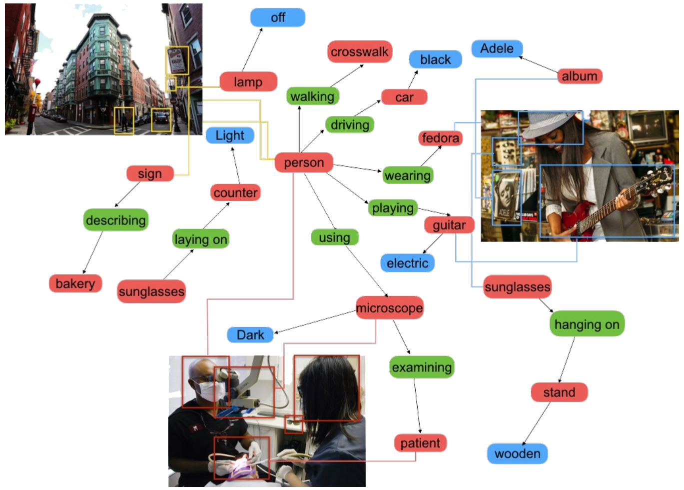
(b) Interpolated contour



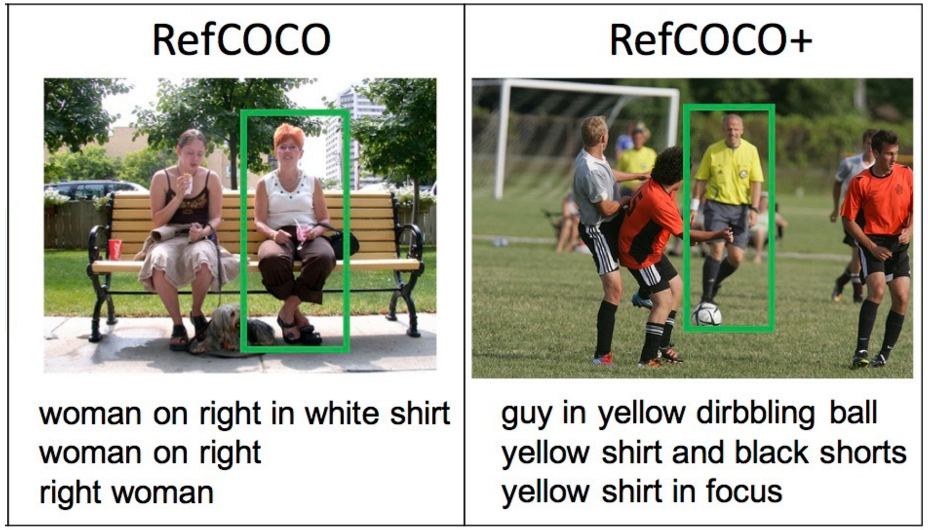
(c) Sampled polygons

Two stage training

- Pre-train on REC task
  - Visual Genome, RefCOCO, RefCOCO+, RefCOCOg datasets, and Flickr entities
  - ~6M distinct language expressions and 164k images in the training set.



A man with pierced ears is wearing glasses and an orange hat.
A man with glasses is wearing a beer can crotched hat.
A man with gauges and glasses is wearing a Blitz hat.
A man in an orange hat starring at something.
A man wears an orange hat and glasses.



Two stage training

- Pre-train on REC task
  - Visual Genome, RefCOCO, RefCOCO+, RefCOCOg datasets, and Flickr entities
  - ~6M distinct language expressions and 164k images in the training set.
- Finetuning on REC + RIS task on RefCOCO, RefCOCO+, RefCOCOg datasets

Referring image segmentation results

	Method	Visual Backbone	Text Encoder	RefCOCO			RefCOCO+			RefCOCOg	
				val	test A	test B	val	test A	test B	val	test
oIoU	STEP [7]	RN101	Bi-LSTM	60.04	63.46	57.97	48.19	52.33	40.41	-	-
	BRINet [29]	RN101	LSTM	60.98	62.99	59.21	48.17	52.32	42.11	-	-
	CMPC [30]	RN101	LSTM	61.36	64.53	59.64	49.56	53.44	43.23	-	-
	LSCM [31]	RN101	LSTM	61.47	64.99	59.55	49.34	53.12	43.50	-	-
	CMPC+ [49]	RN101	LSTM	62.47	65.08	60.82	50.25	54.04	43.47	-	-
	MCN [57]	DN53	Bi-GRU	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40
	EFN [20]	WRN101	Bi-GRU	62.76	65.69	59.67	51.50	55.24	43.01	-	-
	BUSNet [81]	RN101	Self-Att	63.27	66.41	61.39	51.76	56.87	44.13	-	-
	CGAN [56]	DN53	Bi-GRU	64.86	68.04	62.07	51.03	55.51	44.06	51.01	51.69
	LTS [33]	DN53	Bi-GRU	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25
	ReSTR [37]	ViT-B	Transformer	67.22	69.30	64.45	55.78	60.44	48.27	-	-
	PolyFormer-B	Swin-B	BERT-base	74.82	76.64	71.06	67.64	72.89	59.33	67.76	69.05
	PolyFormer-L	Swin-L	BERT-base	**75.96**	**78.29**	**73.25**	**69.33**	**74.56**	**61.87**	**69.20**	**70.19**
mIoU	VLT [19]	DN53	Bi-GRU	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65
	CRIS [76]	RN101	GPT-2	70.47	73.18	66.10	62.27	68.06	53.68	59.87	60.36
	SeqTR [92]	DN53	Bi-GRU	71.70	73.31	69.82	63.04	66.73	58.97	64.69	65.74
	RefTr [42]	RN101	BERT-base	74.34	76.77	70.87	66.75	70.58	59.40	66.63	67.39
	LAVT [84]	Swin-B	BERT-base	74.46	76.89	70.94	65.81	70.97	59.23	63.34	63.62
	PolyFormer-B	Swin-B	BERT-base	75.96	77.09	73.22	70.65	74.51	64.64	69.36	69.88
	PolyFormer-L	Swin-L	BERT-base	**76.94**	**78.49**	**74.83**	**72.15**	**75.71**	**66.73**	**71.15**	**71.17**

Table 1. Comparison with the state-of-the-art methods on three referring image segmentation benchmarks. RN101 denotes ResNet-101 [25], WRN101 refers to Wide ResNet-101 [88], and DN53 denotes Darknet-53 [65].

PolyFormer-B outperforms previous methods on each split of the three datasets

Referring image segmentation results

	Method	Visual Backbone	Text Encoder	RefCOCO			RefCOCO+			RefCOCOg	
				val	test A	test B	val	test A	test B	val	test
oIoU	STEP [7]	RN101	Bi-LSTM	60.04	63.46	57.97	48.19	52.33	40.41	-	-
oIoU	BRINet [29]	RN101	LSTM	60.98	62.99	59.21	48.17	52.32	42.11	-	-
oIoU	CMPC [30]	RN101	LSTM	61.36	64.53	59.64	49.56	53.44	43.23	-	-
oIoU	LSCM [31]	RN101	LSTM	61.47	64.99	59.55	49.34	53.12	43.50	-	-
oIoU	CMPC+ [49]	RN101	LSTM	62.47	65.08	60.82	50.25	54.04	43.47	-	-
oIoU	MCN [57]	DN53	Bi-GRU	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40
oIoU	EFN [20]	WRN101	Bi-GRU	62.76	65.69	59.67	51.50	55.24	43.01	-	-
oIoU	BUSNet [81]	RN101	Self-Att	63.27	66.41	61.39	51.76	56.87	44.13	-	-
oIoU	CGAN [56]	DN53	Bi-GRU	64.86	68.04	62.07	51.03	55.51	44.06	51.01	51.69
oIoU	LTS [33]	DN53	Bi-GRU	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25
oIoU	ReSTR [37]	ViT-B	Transformer	67.22	69.30	64.45	55.78	60.44	48.27	-	-
oIoU	PolyFormer-B	Swin-B	BERT-base	74.82	76.64	71.06	67.64	72.89	59.33	67.76	69.05
oIoU	**PolyFormer-L**	Swin-L	BERT-base	**75.96**	**78.29**	**73.25**	**69.33**	**74.56**	**61.87**	**69.20**	**70.19**
mIoU	VLT [19]	DN53	Bi-GRU	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65
mIoU	CRIS [76]	RN101	GPT-2	70.47	73.18	66.10	62.27	68.06	53.68	59.87	60.36
mIoU	SeqTR [92]	DN53	Bi-GRU	71.70	73.31	69.82	63.04	66.73	58.97	64.69	65.74
mIoU	RefTr [42]	RN101	BERT-base	74.34	76.77	70.87	66.75	70.58	59.40	66.63	67.39
mIoU	LAVT [84]	Swin-B	BERT-base	74.46	76.89	70.94	65.81	70.97	59.23	63.34	63.62
mIoU	PolyFormer-B	Swin-B	BERT-base	75.96	77.09	73.22	70.65	74.51	64.64	69.36	69.88
mIoU	**PolyFormer-L**	Swin-L	BERT-base	**76.94**	**78.49**	**74.83**	**72.15**	**75.71**	**66.73**	**71.15**	**71.17**

Table 1. Comparison with the state-of-the-art methods on three referring image segmentation benchmarks. RN101 denotes ResN
101 [25], WRN101 refers to Wide ResNet-101 [88], and DN53 denotes Darknet-53 [65].

+3.9%, 3.93%, 5.24% mIoU on challenging RefCOCO+

Referring image segmentation results

	Method	Visual Backbone	Text Encoder	RefCOCO			RefCOCO+			RefCOCOg	
				val	test A	test B	val	test A	test B	val	test
oIoU	STEP [7]	RN101	Bi-LSTM	60.04	63.46	57.97	48.19	52.33	40.41	-	-
	BRINet [29]	RN101	LSTM	60.98	62.99	59.21	48.17	52.32	42.11	-	-
	CMPC [30]	RN101	LSTM	61.36	64.53	59.64	49.56	53.44	43.23	-	-
	LSCM [31]	RN101	LSTM	61.47	64.99	59.55	49.34	53.12	43.50	-	-
	CMPC+ [49]	RN101	LSTM	62.47	65.08	60.82	50.25	54.04	43.47	-	-
	MCN [57]	DN53	Bi-GRU	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40
	EFN [20]	WRN101	Bi-GRU	62.76	65.69	59.67	51.50	55.24	43.01	-	-
	BUSNet [81]	RN101	Self-Att	63.27	66.41	61.39	51.76	56.87	44.13	-	-
	CGAN [56]	DN53	Bi-GRU	64.86	68.04	62.07	51.03	55.51	44.06	51.01	51.69
	LTS [33]	DN53	Bi-GRU	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25
	ReSTR [37]	ViT-B	Transformer	67.22	69.30	64.45	55.78	60.44	48.27	-	-
	PolyFormer-B	Swin-B	BERT-base	74.82	76.64	71.06	67.64	72.89	59.33	67.76	69.05
	PolyFormer-L	Swin-L	BERT-base	**75.96**	**78.29**	**73.25**	**69.33**	**74.56**	**61.87**	**69.20**	**70.19**
mIoU	VLT [19]	DN53	Bi-GRU	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65
	CRIS [76]	RN101	GPT-2	70.47	73.18	66.10	62.27	68.06	53.68	59.87	60.36
	SeqTR [92]	DN53	Bi-GRU	71.70	73.31	69.82	63.04	66.73	58.97	64.69	65.74
	RefTr [42]	RN101	BERT-base	74.34	76.77	70.87	66.75	70.58	59.40	66.63	67.39
	LAVT [84]	Swin-B	BERT-base	74.46	76.89	70.94	65.81	70.97	59.23	63.34	63.62
	PolyFormer-B	Swin-B	BERT-base	75.96	77.09	73.22	70.65	74.51	64.64	69.36	69.88
	PolyFormer-L	Swin-L	BERT-base	**76.94**	**78.49**	**74.83**	**72.15**	**75.71**	**66.73**	**71.15**	**71.17**

Table 1. Comparison with the state-of-the-art methods on three referring image segmentation benchmarks. RN101 denotes ResNet-101 [25], WRN101 refers to Wide ResNet-101 [88], and DN53 denotes Darknet-53 [65].

+2.73%, 2.49% mIoU on most challenging RefCOCOg

Referring image segmentation results

	Method	Visual Backbone	Text Encoder	RefCOCO			RefCOCO+			RefCOCOg	
				val	test A	test B	val	test A	test B	val	test
oIoU	STEP [7]	RN101	Bi-LSTM	60.04	63.46	57.97	48.19	52.33	40.41	-	-
	BRINet [29]	RN101	LSTM	60.98	62.99	59.21	48.17	52.32	42.11	-	-
	CMPC [30]	RN101	LSTM	61.36	64.53	59.64	49.56	53.44	43.23	-	-
	LSCM [31]	RN101	LSTM	61.47	64.99	59.55	49.34	53.12	43.50	-	-
	CMPC+ [49]	RN101	LSTM	62.47	65.08	60.82	50.25	54.04	43.47	-	-
	MCN [57]	DN53	Bi-GRU	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40
	EFN [20]	WRN101	Bi-GRU	62.76	65.69	59.67	51.50	55.24	43.01	-	-
	BUSNet [81]	RN101	Self-Att	63.27	66.41	61.39	51.76	56.87	44.13	-	-
	CGAN [56]	DN53	Bi-GRU	64.86	68.04	62.07	51.03	55.51	44.06	51.01	51.69
	LTS [33]	DN53	Bi-GRU	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25
	ReSTR [37]	ViT-B	Transformer	67.22	69.30	64.45	55.78	60.44	48.27	-	-
	PolyFormer-B	Swin-B	BERT-base	74.82	76.64	71.06	67.64	72.89	59.33	67.76	69.05
	PolyFormer-L	Swin-L	BERT-base	**75.96**	**78.29**	**73.25**	**69.33**	**74.56**	**61.87**	**69.20**	**70.19**
mIoU	VLT [19]	DN53	Bi-GRU	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65
	CRIS [76]	RN101	GPT-2	70.47	73.18	66.10	62.27	68.06	53.68	59.87	60.36
	SeqTR [92]	DN53	Bi-GRU	71.70	73.31	69.82	63.04	66.73	58.97	64.69	65.74
	RefTr [42]	RN101	BERT-base	74.34	76.77	70.87	66.75	70.58	59.40	66.63	67.39
	LAVT [84]	Swin-B	BERT-base	74.46	76.89	70.94	65.81	70.97	59.23	63.34	63.62
	PolyFormer-B	Swin-B	BERT-base	75.96	77.09	73.22	70.65	74.51	64.64	69.36	69.88
	PolyFormer-L	Swin-L	BERT-base	**76.94**	**78.49**	**74.83**	**72.15**	**75.71**	**66.73**	**71.15**	**71.17**

Table 1. Comparison with the state-of-the-art methods on three referring image segmentation benchmarks. RN101 denotes ResNet-101 [25], WRN101 refers to Wide ResNet-101 [88], and DN53 denotes Darknet-53 [65].

PolyFormer-L vs. B:
+1~2 points

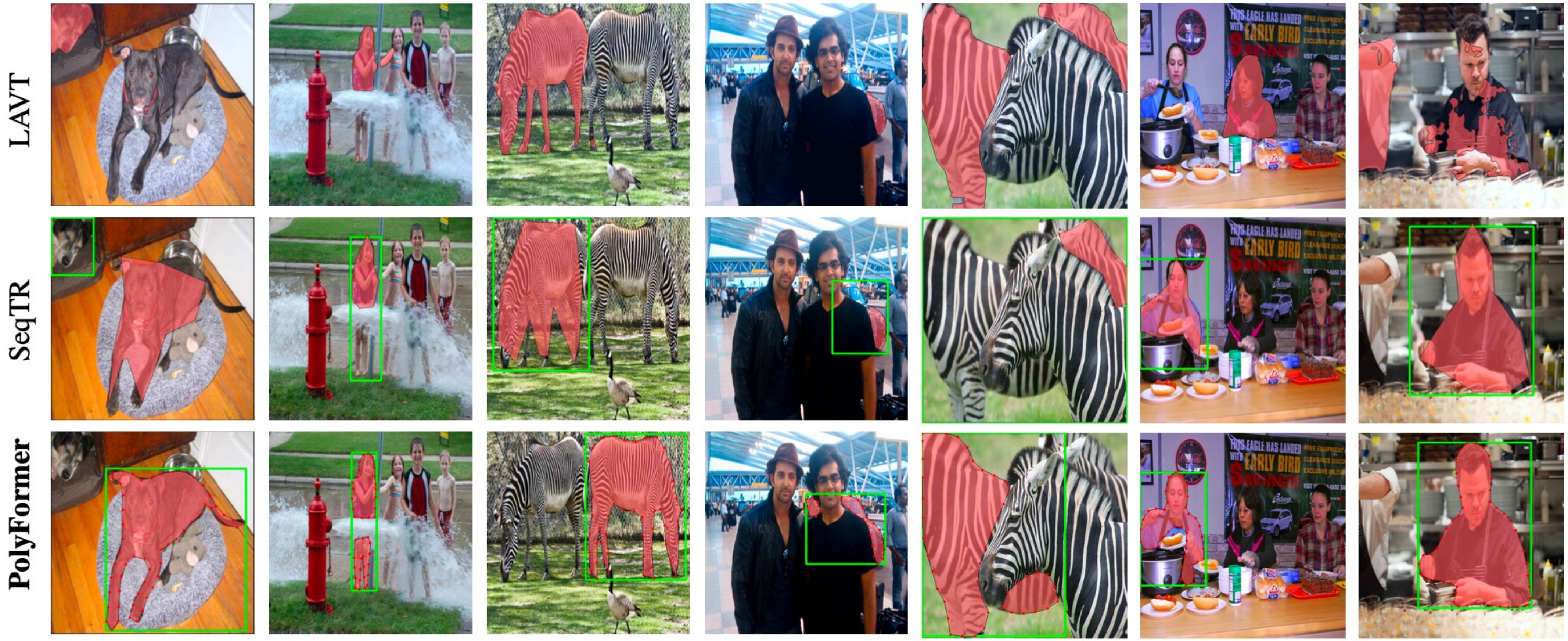
Zero-shot Transfer to Referring Video Object Segmentation

Method	Visual Backbone	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
CMSA+RNN [85]	ResNet-50	40.2	36.9	43.5
URVOS [70]	ResNet-50	51.5	47.3	56.0
CITD [44]	ResNet-101	56.4	54.8	58.1
ReferFormer [78]	Swin-L	60.5	57.6	63.4
ReferFormer [78]	Video-Swin-B	61.1	**58.1**	64.1
PolyFormer-B†	Swin-B	60.9	56.6	65.2
PolyFormer-L†	Swin-L	**61.5**	57.2	**65.8**

Table 3. Comparison with the state-of-the-art methods on Ref-DAVIS17. †means our model is trained on image datasets only. ReferFormer is trained on both image and video datasets.

Best J&F w/o training on video

Visualization Results on RefCOCOg



(a) "a dark grey dog on a light grey round bed wearing a red collar"

(b) "girl in purple"

(c) "zebra eating grass with a goose in front of it"

(d) "a black car parked at a transportation terminal"

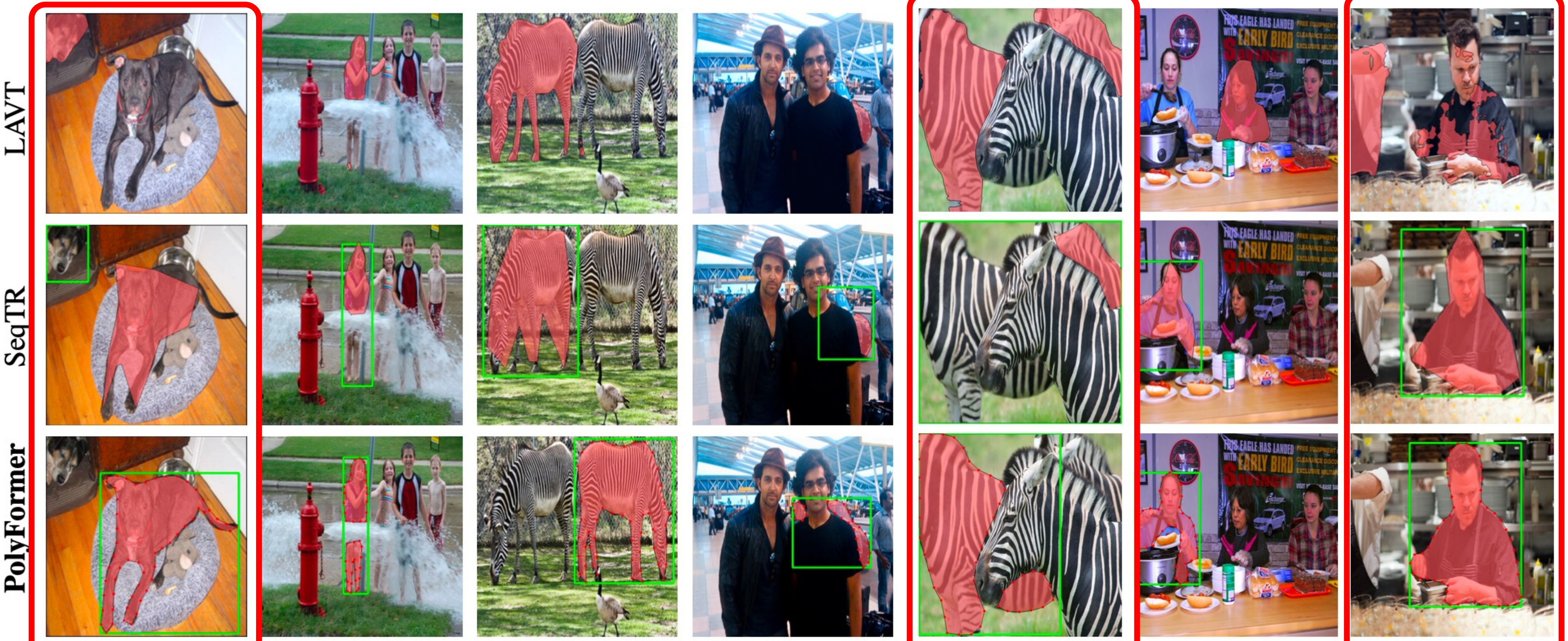
(e) "a zebra with its head not visible but much of its body able to be seen"

(f) " a girl was cooking the food and serving"

(g) "a man wearing a black shirt and a black and white striped apron stirring something in a metal container"

Visualization Results on RefCOCOg

complex language understanding



(a) "a dark grey dog on a light grey round bed wearing a red collar"
(b) "girl in purple"
(c) "zebra eating grass with a goose in front of it"
(d) "a black car parked at a transportation terminal"
(e) "a zebra with its head not visible but much of its body able to be seen"
f) " a girl was cooking he food and serving"
(g) "a man wearing a black shirt and a black and white striped apron stirring something in a metal container"

Visualization Results on RefCOCOg

(a) "a dark grey dog on a light grey round bed wearing a red collar"

(b) "girl in purple"

(c) "zebra eating grass with a goose in front of it"

(d) "a black car parked at a transportation terminal"

(e) "a zebra with its head not visible but much of its body able to be seen"

(f) " a girl was cooking the food and serving"

(g) "a man wearing a black shirt and a black and white striped apron stirring something in a metal container"

Visualization Results on RefCOCOg

complex vision-language semantics

(a) "a dark grey dog on a light grey round bed wearing a red collar" (b) "girl in purple" (c) "zebra eating grass with a goose in front of it" (d) "a black car parked at a transportation terminal" (e) "a zebra with its head not visible but much of its body able to be seen" (f) " a girl was cooking the food and serving" (g) "a man wearing a black shirt and a black and white striped apron stirring something in a metal container"

Zero-shot Evaluation on Stable Diffusion Images

(a) "A cat chef cooking fish in a fancy restaurant"

(b) "A chair that looks like octopus"

(c) "A small cabin on top of a snowy mountain in the style of Disney artstation"

(d) "A shiba inu puppy painted by Monet"

(e) "A gentleman otter in a 19th century portrait"

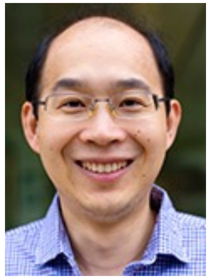
(f) "A pikachu fine-dining with a view to the Eiffel Tower"

(g) "A pig robot preparing a delicious meal"

Paper tag: THU-AM-208

PolyFormer: Referring Image Segmentation as Sequential Polygon Generation




Jiang Liu[1*†] Hui Ding[2*] Zhaowei Cai[2] Yuting Zhang[2] Ravi Kumar Satzoda[2] Vijay Mahadevan[2] R. Manmatha[2]

[1] Johns Hopkins University, [2] AWS AI Labs, [*]Equal Contribution, [†]Work done during internship at AWS AI Labs

https://polyformer.github.io/

Code & Demo Available