code

Paper

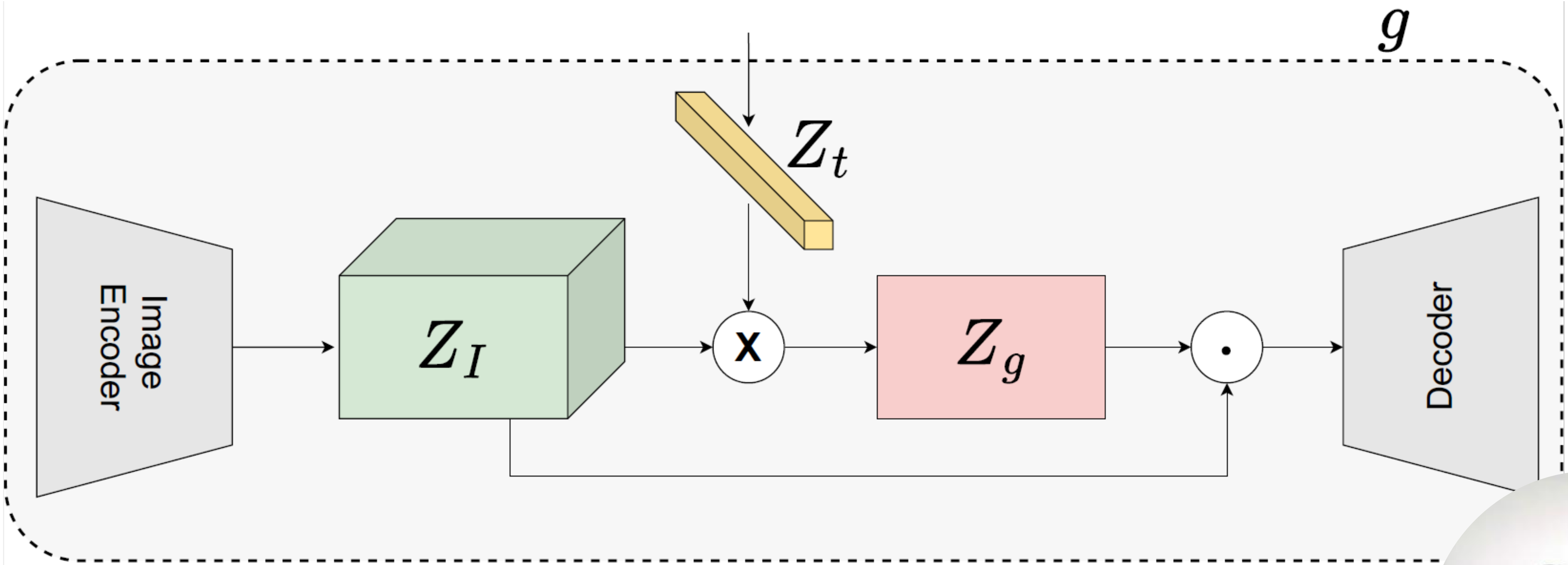
TEL AVIV UNIVERSITY

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

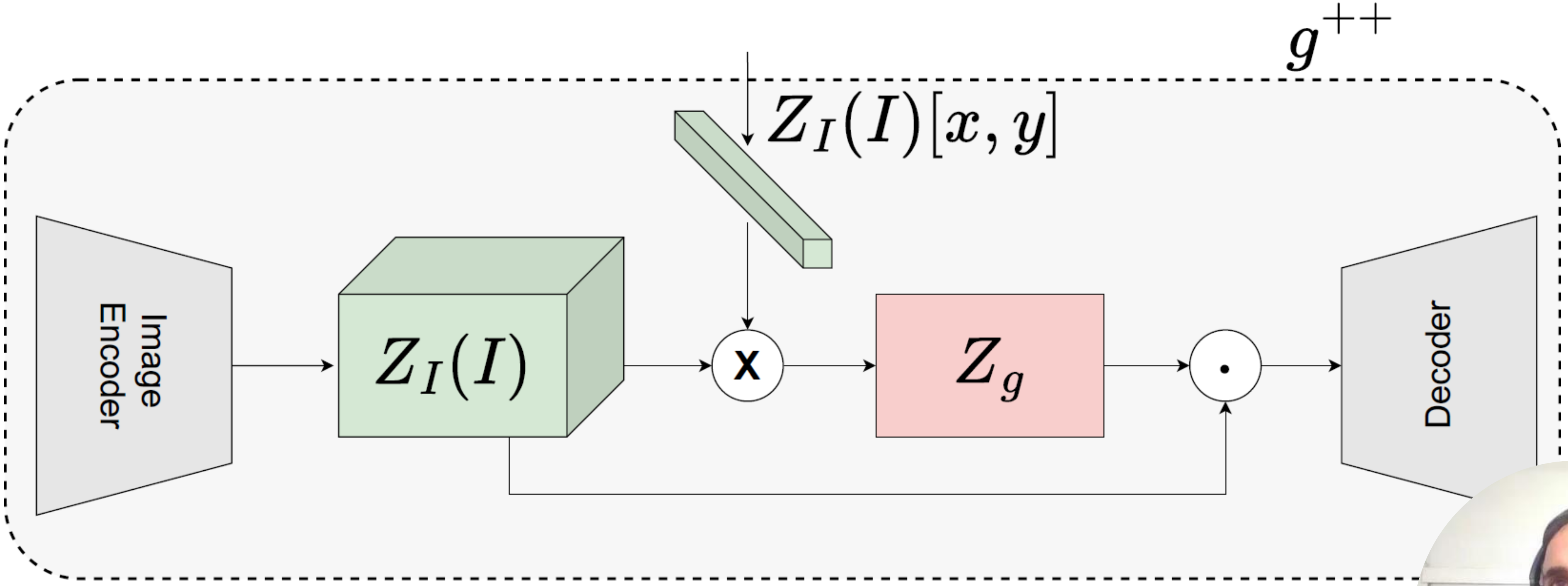
Similarity Maps for Self-Training Weakly-Supervised Phrase Grounding

Tal Shaharabany, Lior Wolf

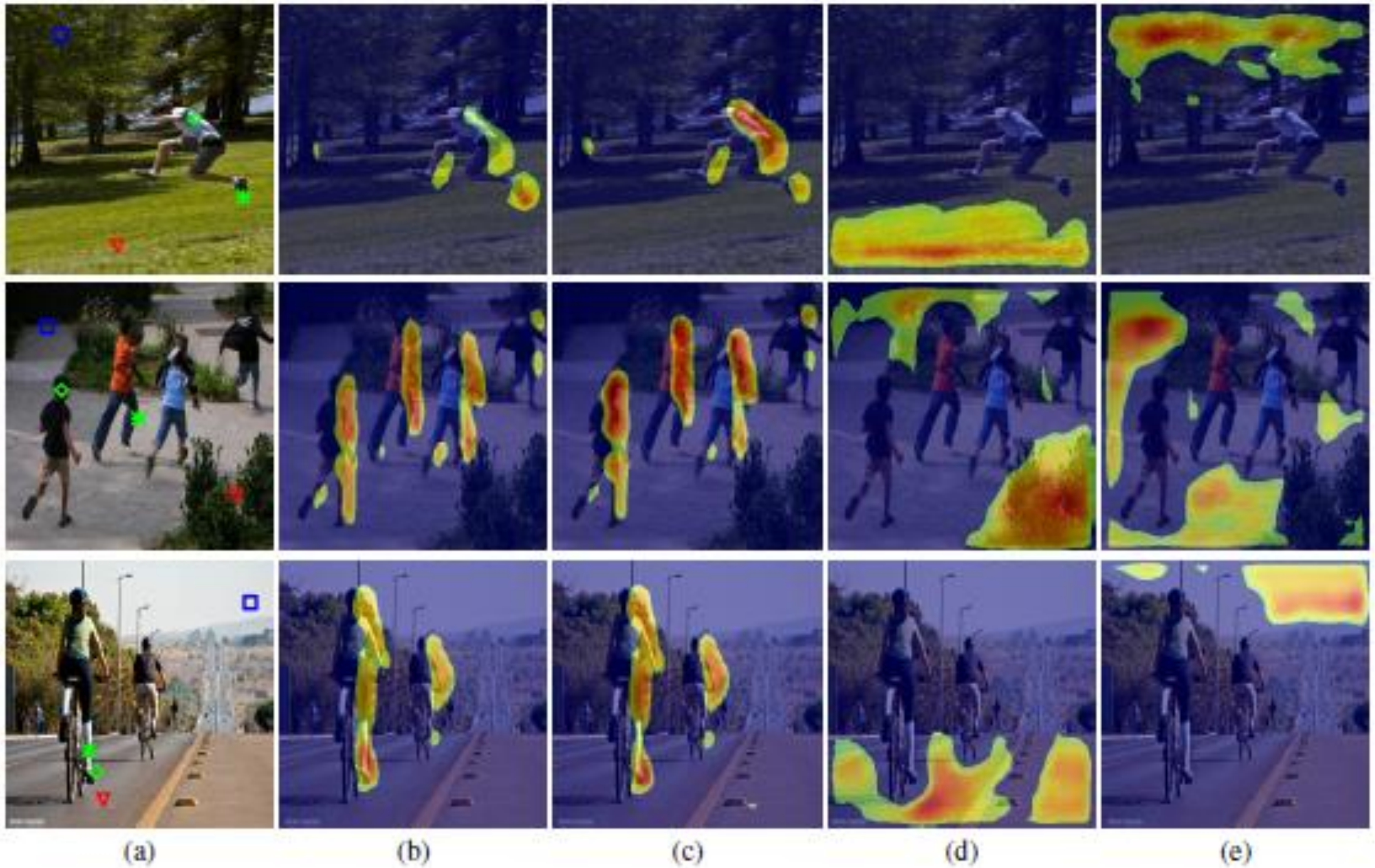
Method – Self Similarity Maps



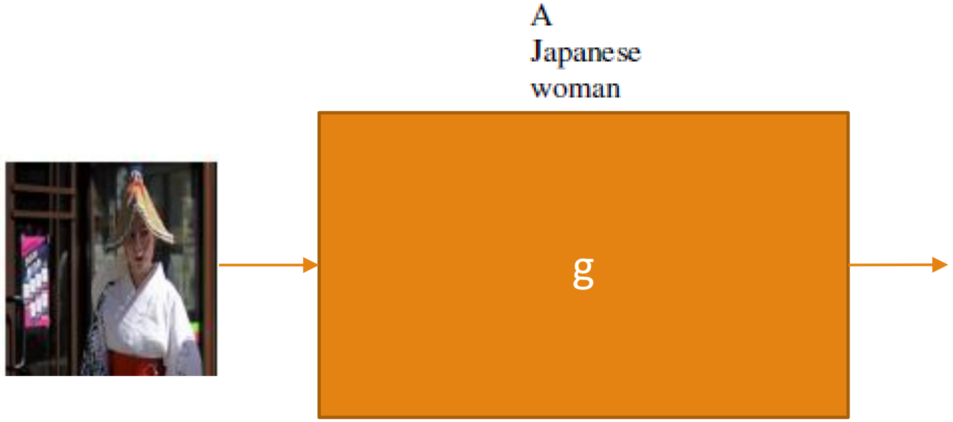
Method – Self Similarity Maps



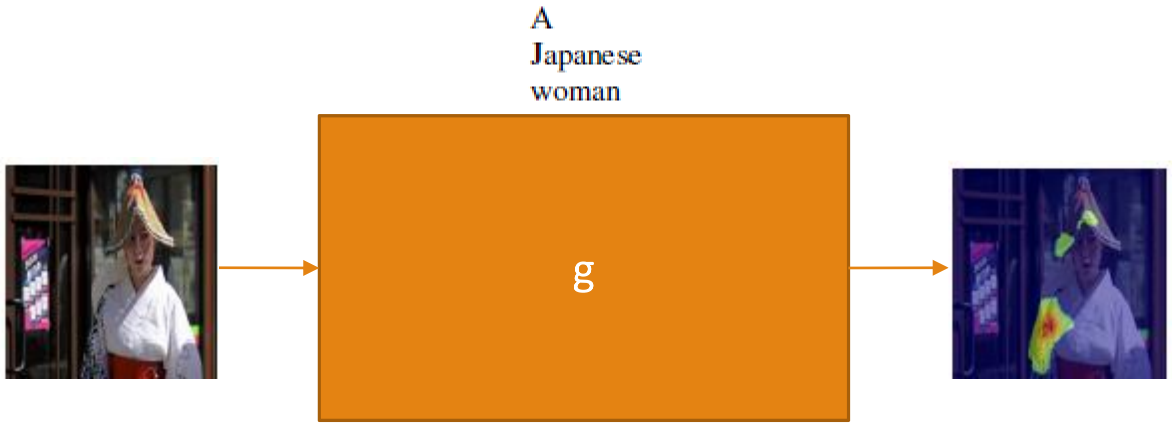
Method – Self Similarity Maps



Method – Maps Selection



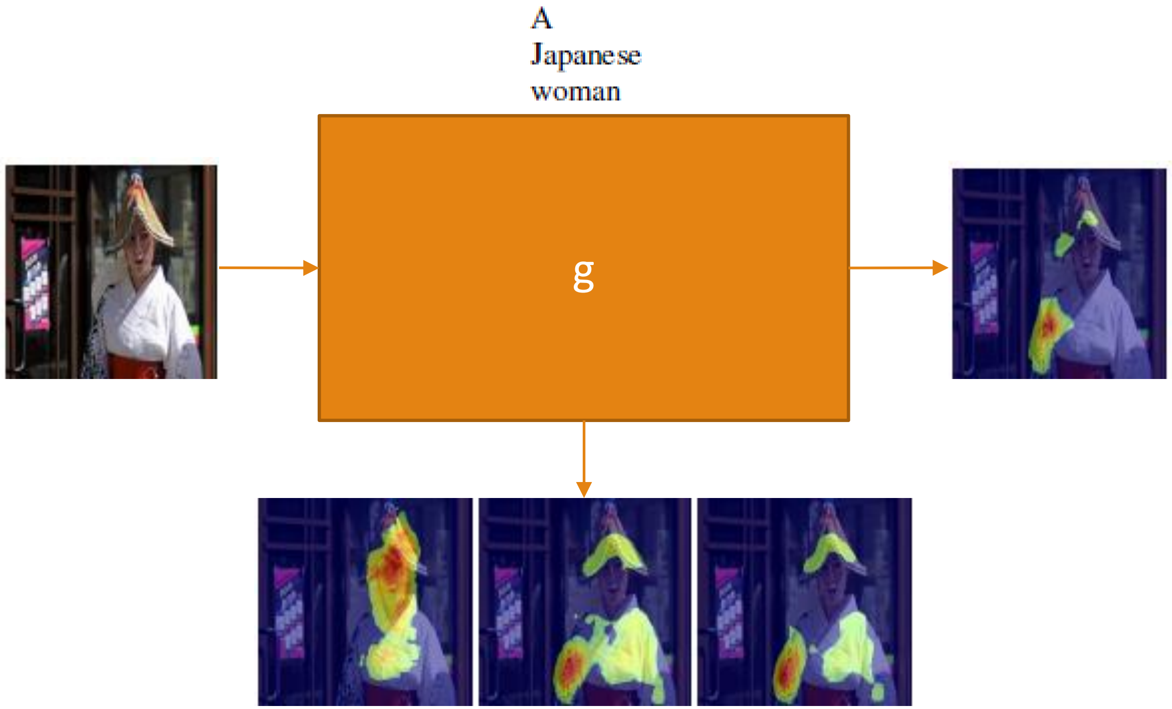
Method – Maps Selection



A Japanese woman

g

Method – Maps Selection

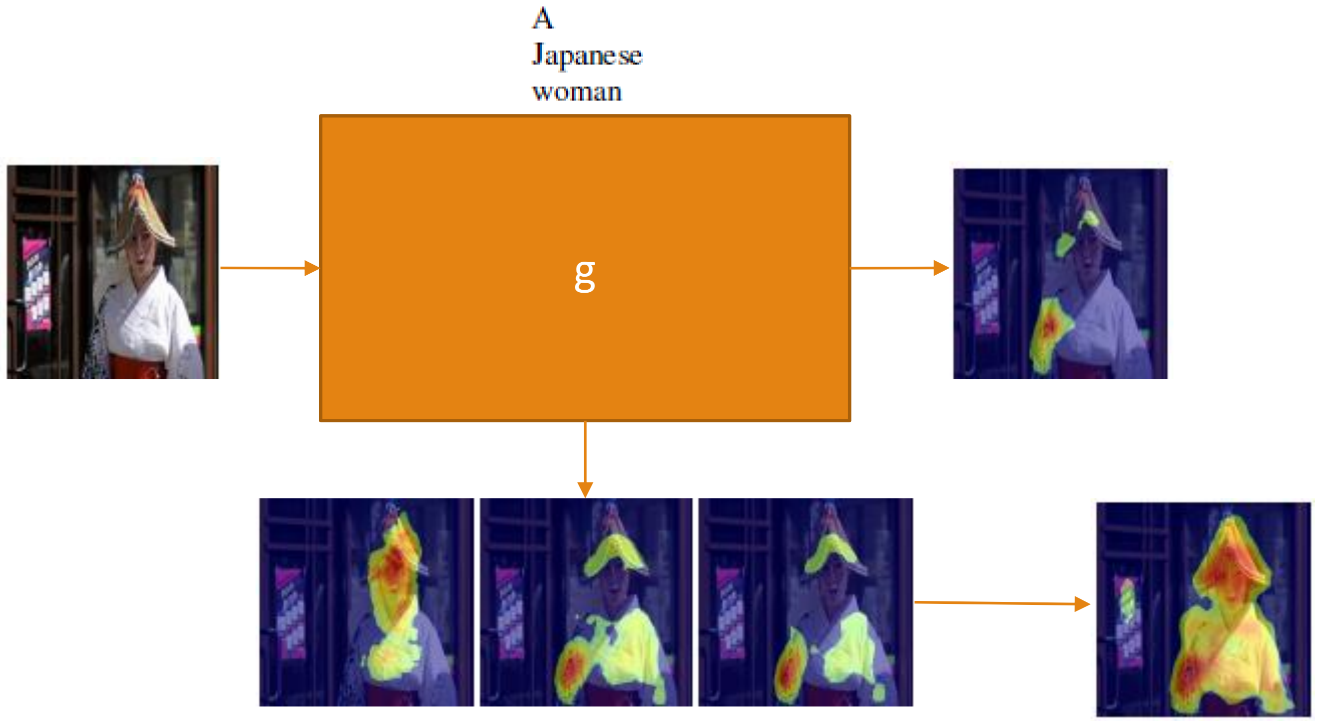


A Japanese woman

g

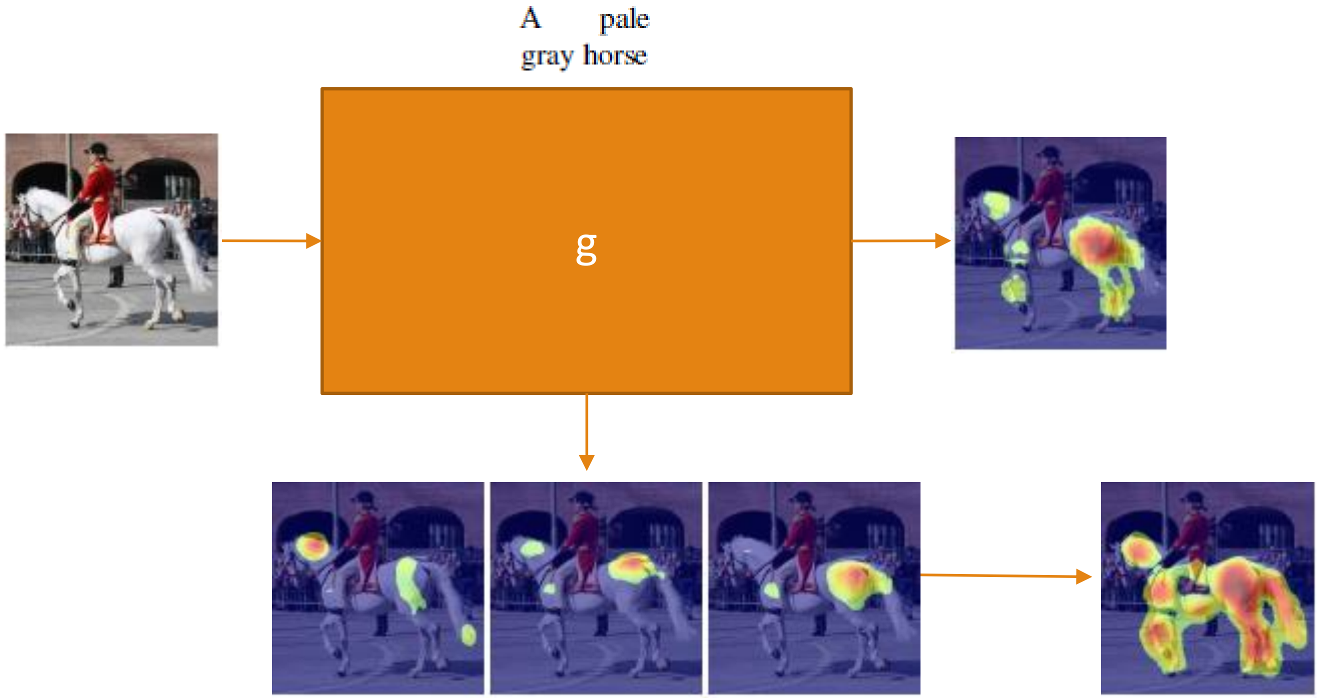
K relevance maps

Method – Maps Selection



K relevance maps

Method – Maps Selection

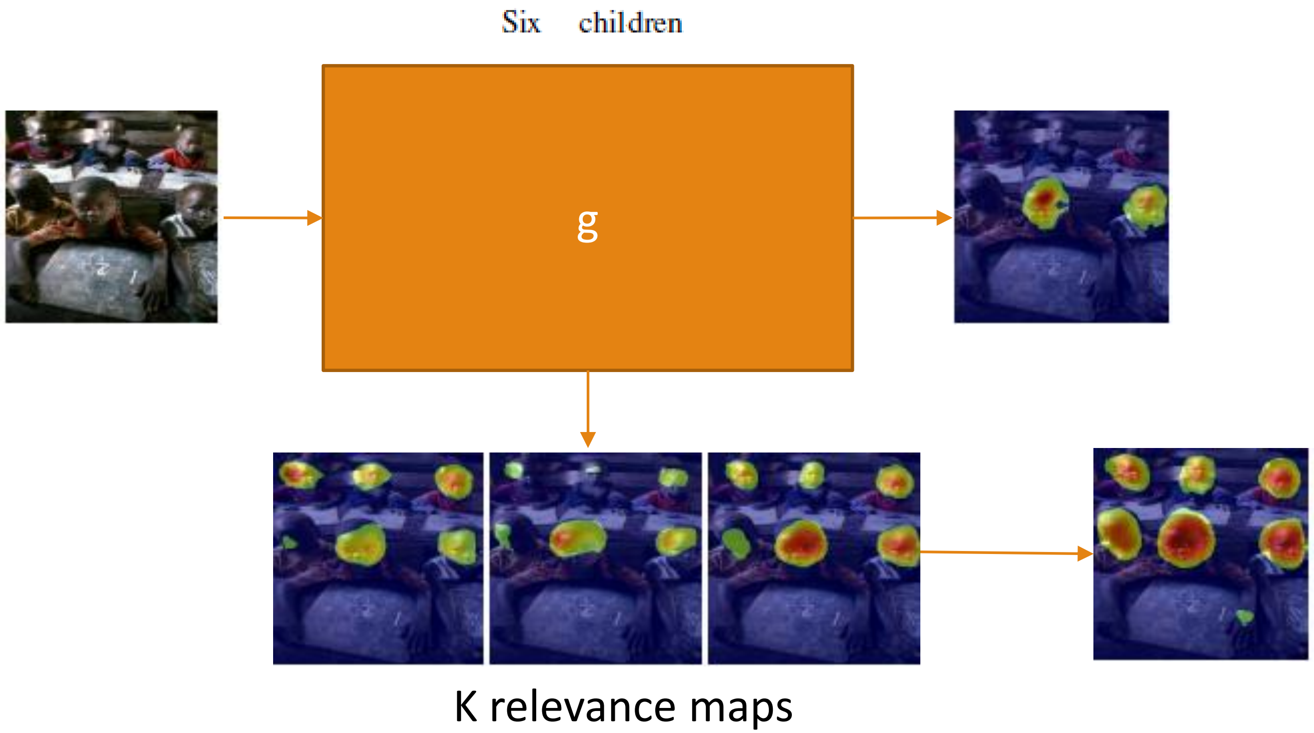


A pale gray horse

g

K relevance maps

Method – Maps Selection

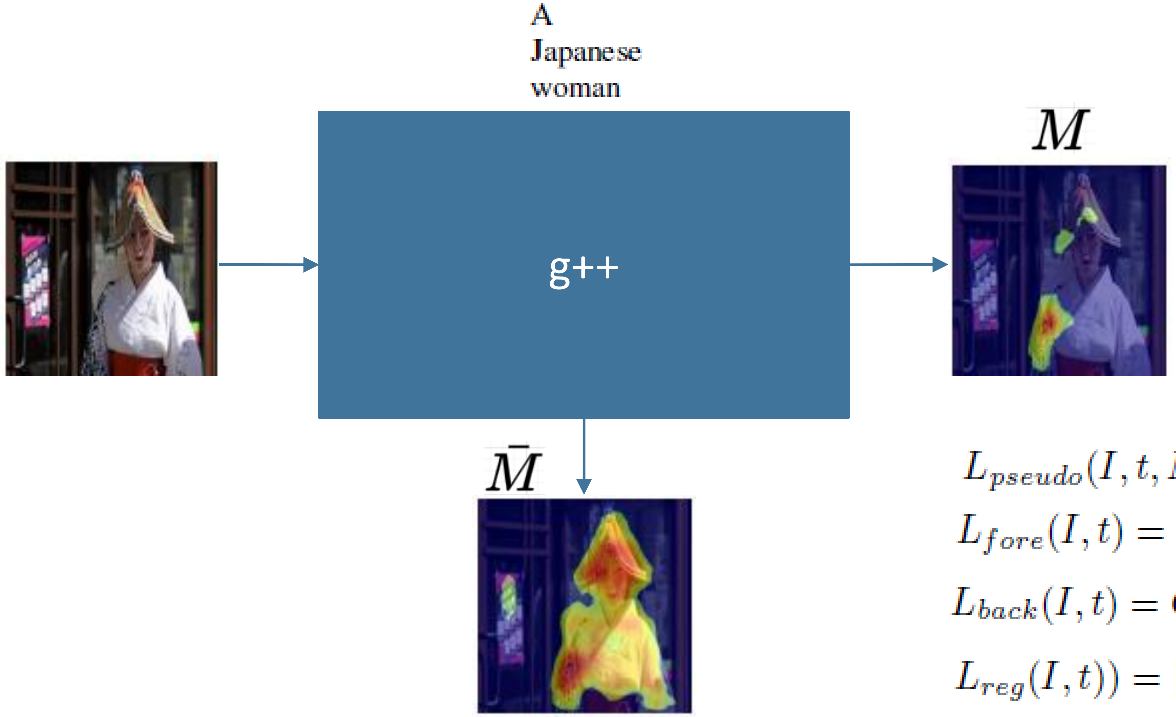


Six children

g

K relevance maps

Method – Fine-tune



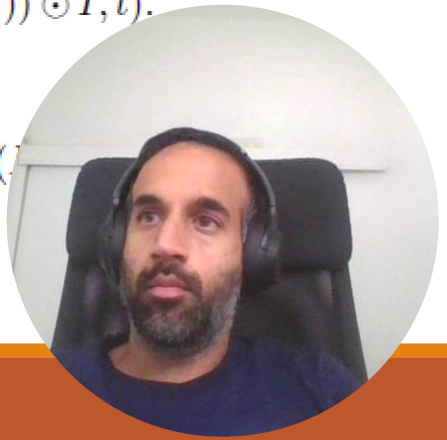
$$L_{pseudo}(I,t,\bar{M}) = \|\bar{M} - g^{++}(I,Z_t(t))\|^2 ,$$

$$L_{fore}(I,t) = -CLIP(g^{++}(I,Z_t(t)) \odot I,t),$$

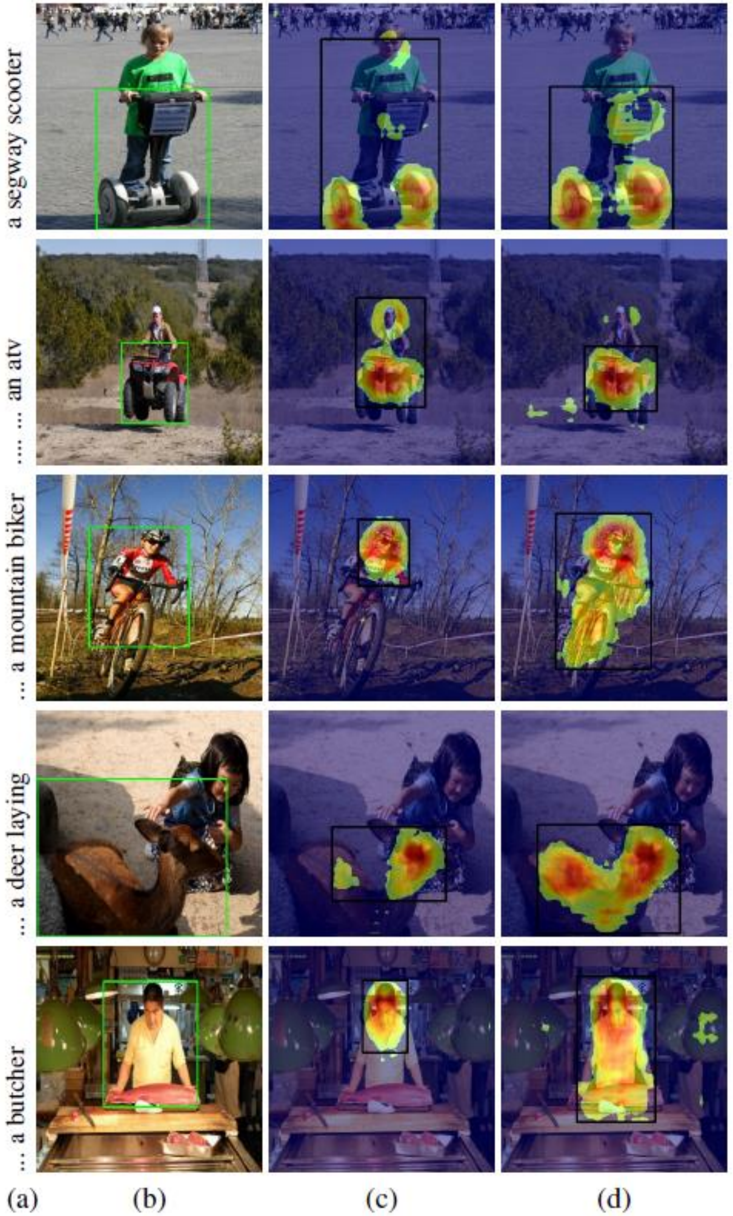
$$L_{back}(I,t) = CLIP((1 - g^{++}(I,Z_t(t))) \odot I,t).$$

$$L_{reg}(I,t)) = \|g^{++}(I,Z_t(t))\|$$

$$L(I,t,\bar{M}) \quad = \quad L_{pseudo}(I,t,\bar{M}) \; + \; L_{fore}(I,t) \; + L_{back}($$

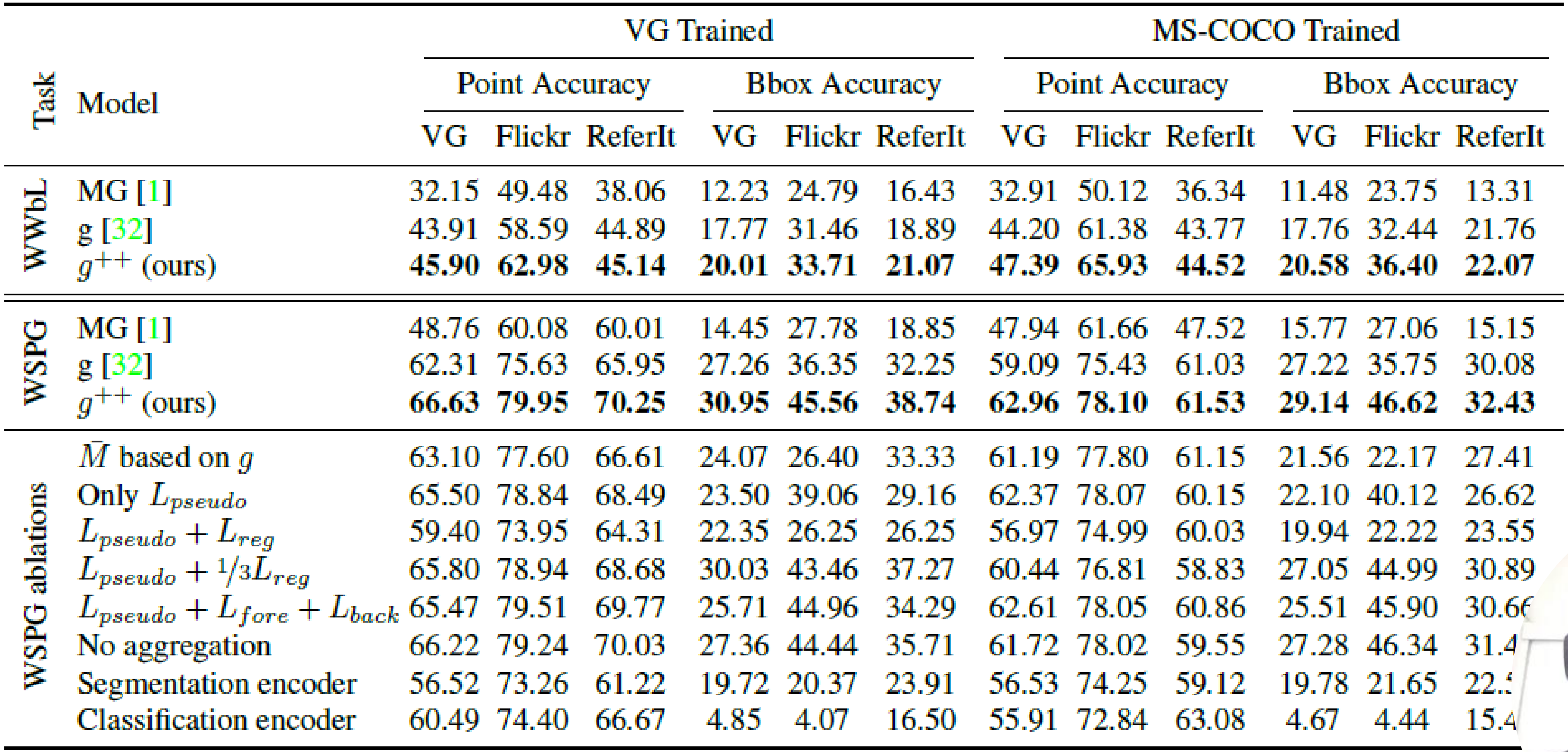


Fine-tune Visualization



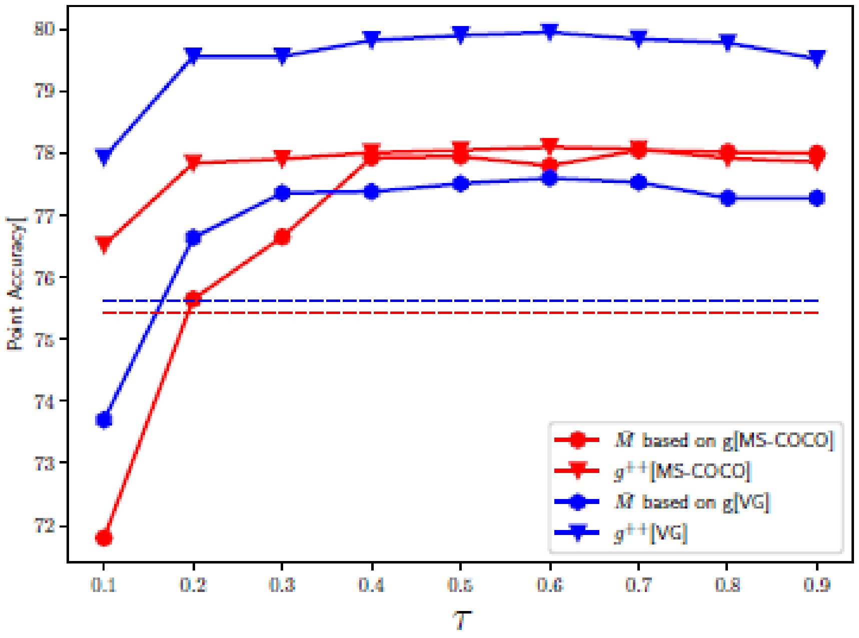
(a) Input text (b) input image + bbox gt (c) g output (d) g++ output

Results and Ablation Study

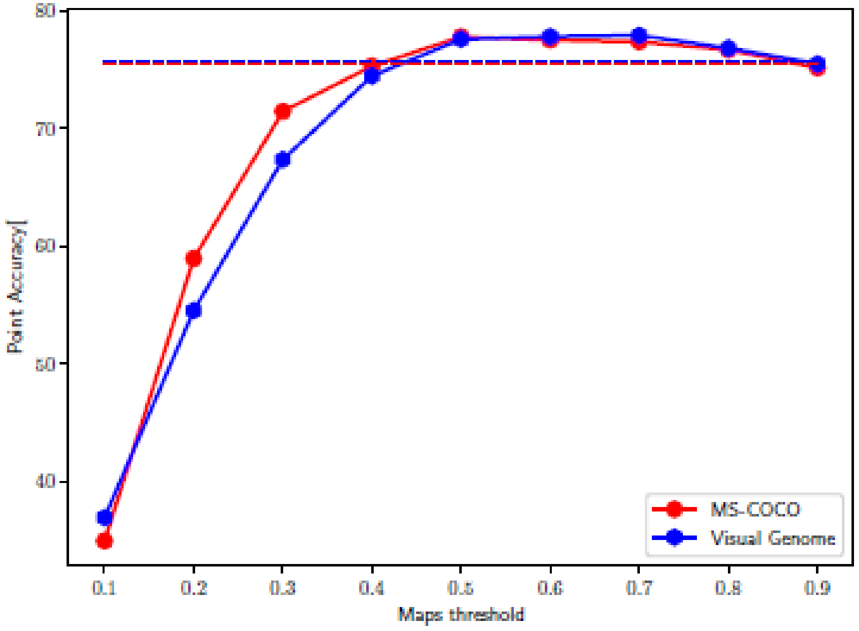


Task	Model	VG Trained						MS-COCO Trained					
		Point Accuracy			Bbox Accuracy			Point Accuracy			Bbox Accuracy		
		VG	Flickr	ReferIt	VG	Flickr	ReferIt	VG	Flickr	ReferIt	VG	Flickr	ReferIt
WWbL	MG [1]	32.15	49.48	38.06	12.23	24.79	16.43	32.91	50.12	36.34	11.48	23.75	13.31
	g [32]	43.91	58.59	44.89	17.77	31.46	18.89	44.20	61.38	43.77	17.76	32.44	21.76
	g^{++} (ours)	**45.90**	**62.98**	**45.14**	**20.01**	**33.71**	**21.07**	**47.39**	**65.93**	**44.52**	**20.58**	**36.40**	**22.07**
WSPG	MG [1]	48.76	60.08	60.01	14.45	27.78	18.85	47.94	61.66	47.52	15.77	27.06	15.15
	g [32]	62.31	75.63	65.95	27.26	36.35	32.25	59.09	75.43	61.03	27.22	35.75	30.08
	g^{++} (ours)	**66.63**	**79.95**	**70.25**	**30.95**	**45.56**	**38.74**	**62.96**	**78.10**	**61.53**	**29.14**	**46.62**	**32.43**
WSPG ablations	$\bar{M}$ based on g	63.10	77.60	66.61	24.07	26.40	33.33	61.19	77.80	61.15	21.56	22.17	27.41
	Only L_{pseudo}	65.50	78.84	68.49	23.50	39.06	29.16	62.37	78.07	60.15	22.10	40.12	26.62
	$L_{pseudo} + L_{reg}$	59.40	73.95	64.31	22.35	26.25	26.25	56.97	74.99	60.03	19.94	22.22	23.55
	$L_{pseudo} + {}^{1}/{}_{3}L_{reg}$	65.80	78.94	68.68	30.03	43.46	37.27	60.44	76.81	58.83	27.05	44.99	30.89
	$L_{pseudo} + L_{fore} + L_{back}$	65.47	79.51	69.77	25.71	44.96	34.29	62.61	78.05	60.86	25.51	45.90	30.6
	No aggregation	66.22	79.24	70.03	27.36	44.44	35.71	61.72	78.02	59.55	27.28	46.34	31.4
	Segmentation encoder	56.52	73.26	61.22	19.72	20.37	23.91	56.53	74.25	59.12	19.78	21.65	22.
	Classification encoder	60.49	74.40	66.67	4.85	4.07	16.50	55.91	72.84	63.08	4.67	4.44	15.4

Parameter Sensitivity Analysis

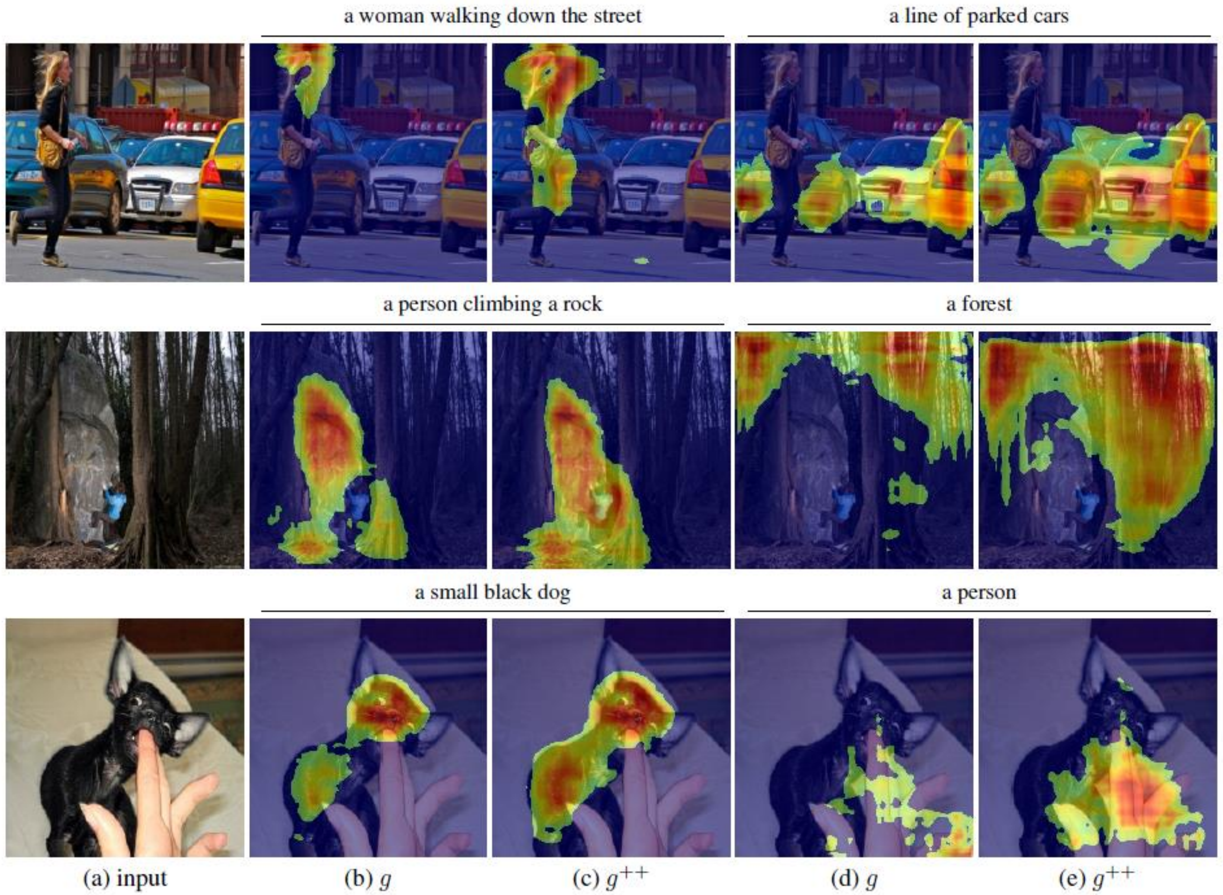


(a)



(b)

WWbL Visualization with g++



code

Paper

Similarity Maps for Self-Training Weakly-Supervised Phrase Grounding – CVPR 23’

Tal Shaharabany, Lior Wolf