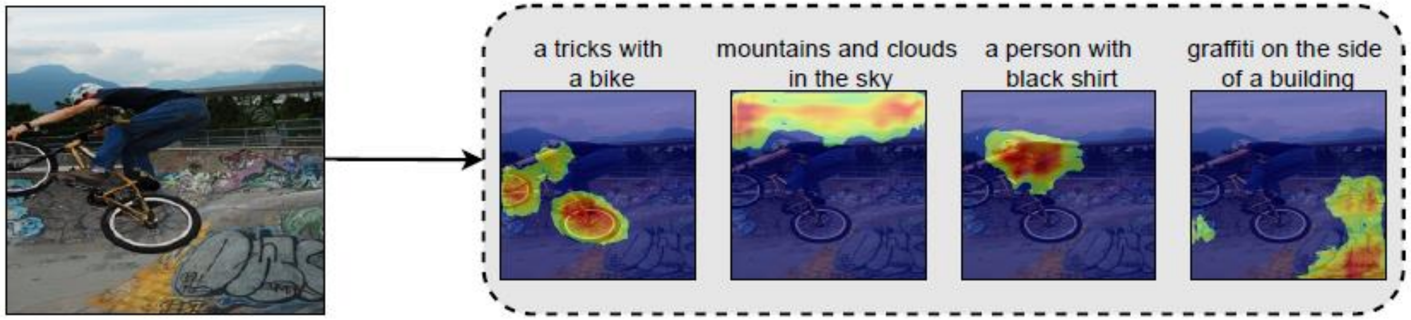
TEL AVIV UNIVERSITY


a tricks with a bike
mountains and clouds in the sky
a person with black shirt
graffiti on the side of a building

Open-Vocabulary Weakly Supervised Visual Recognition Algorithms

Tal Shaharabany

Under the Supervision of
Prof. Lior Wolf

What does it mean, to see? The plain man’s answer (and Aristotle’s too) would be, to know what is
where by looking. In other words, vision is the process of discovering from images what is present in
the world, and where it is.

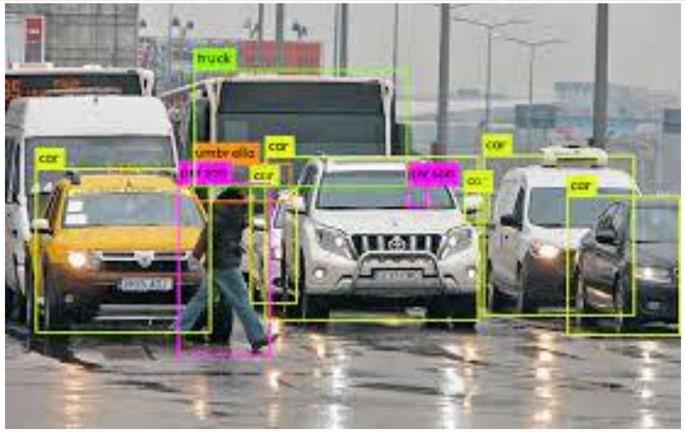
- (1982) Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. San Francisco: W. H. Freeman and Company.

Landscape

Name	Open-world	Weakly supervised	Purely visual
Object Detection	X	X	V
Phrase-Grounding	V	X	X
Weakly supervised localization	X	V	V
Weakly supervised Phrase-Grounding	V	V	X
What is where by looking (WWbL)	V	V	V

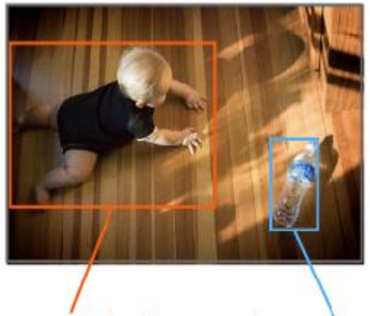
Object Detection

Name	Open-world	Weakly supervised	Purely visual
Object Detection	**X**	**X**	**V**
Phrase-Grounding	V	X	X
Weakly supervised localization	X	V	V
Weakly supervised Phrase-Grounding	V	V	X
What is where by looking (WWbL)	V	V	V



Phrase-Grounding

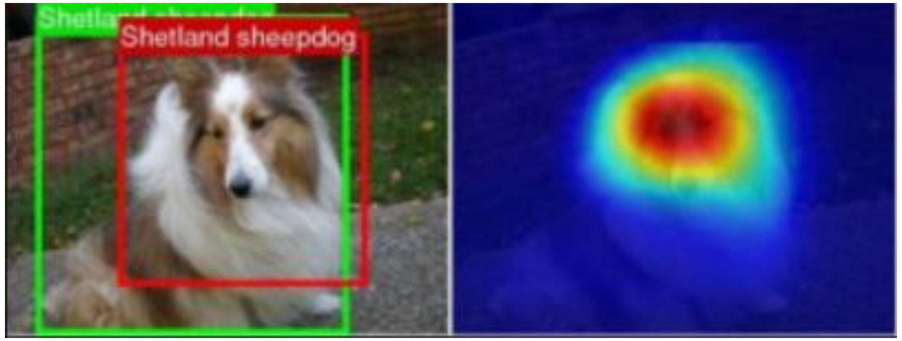
Name	Open-world	Weakly supervised	Purely visual
Object Detection	X	X	V
Phrase-Grounding	**V**	**X**	**X**
Weakly supervised localization	X	V	V
Weakly supervised Phrase-Grounding	V	V	X
What is where by looking (WWbL)	V	V	V



A young baby crawls across the wood floor towards the water bottle

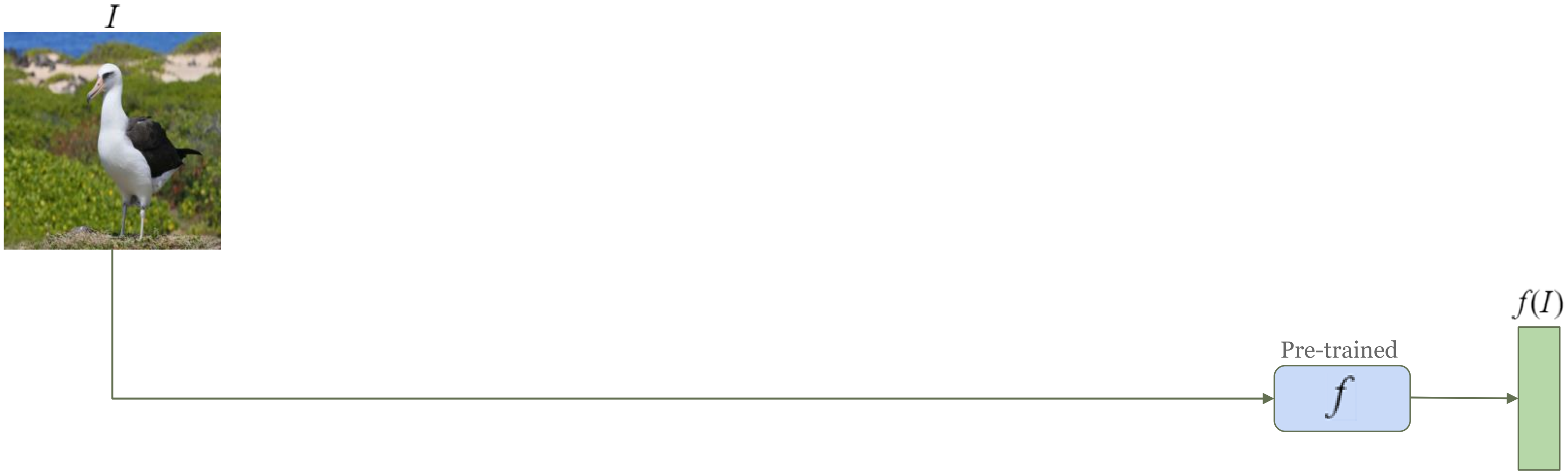
Weakly supervised localization

Name	Open-world	Weakly supervised	Purely visual
Object Detection	X	X	V
Phrase-Grounding	V	X	X
Weakly supervised localization	**X**	**V**	**V**
Weakly supervised Phrase-Grounding	V	V	X
What is where by looking (WWbL)	V	V	V



Learning a Weight Map

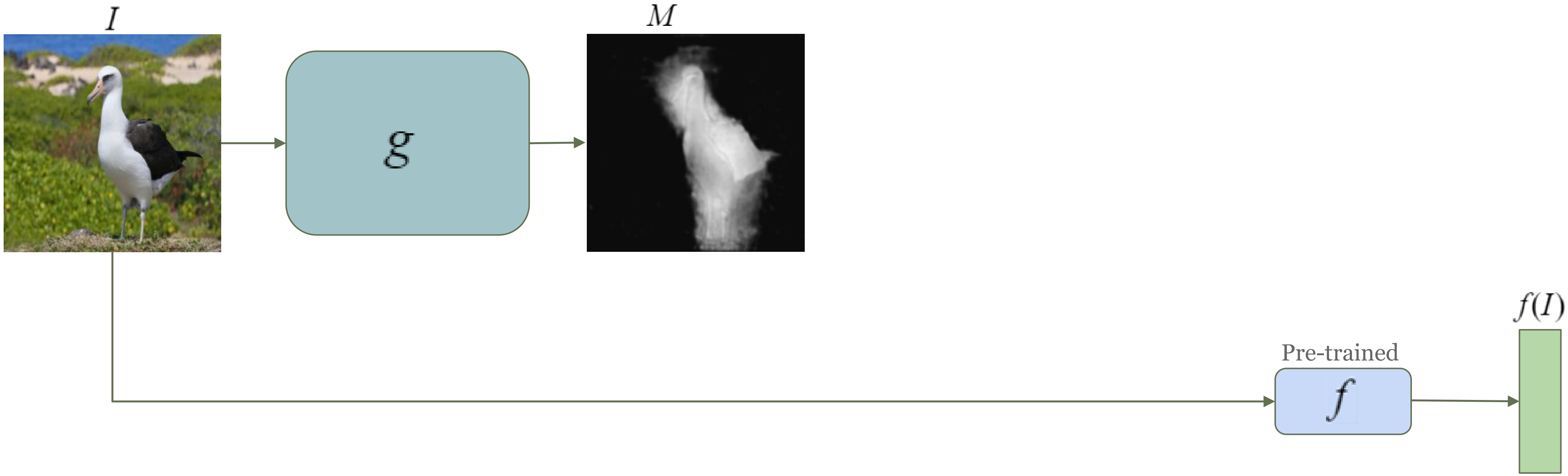
A generic approach that does not assume anything on the AI model



T. Shaharabany, L. Wolf. **Learning a Weight Map for Weakly-Supervised Localization.** ICASSP 23'

Learning a Weight Map

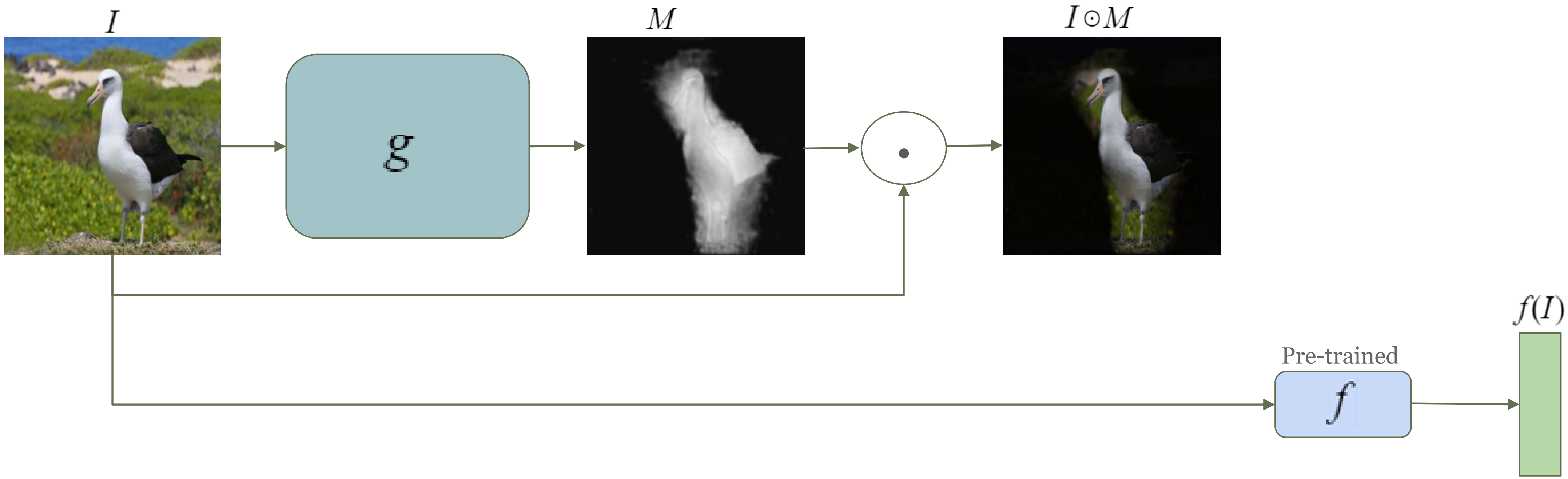
A generic approach that does not assume anything on the AI model



T. Shaharabany, L. Wolf. **Learning a Weight Map for Weakly-Supervised Localization.** ICASSP 23’

Learning a Weight Map

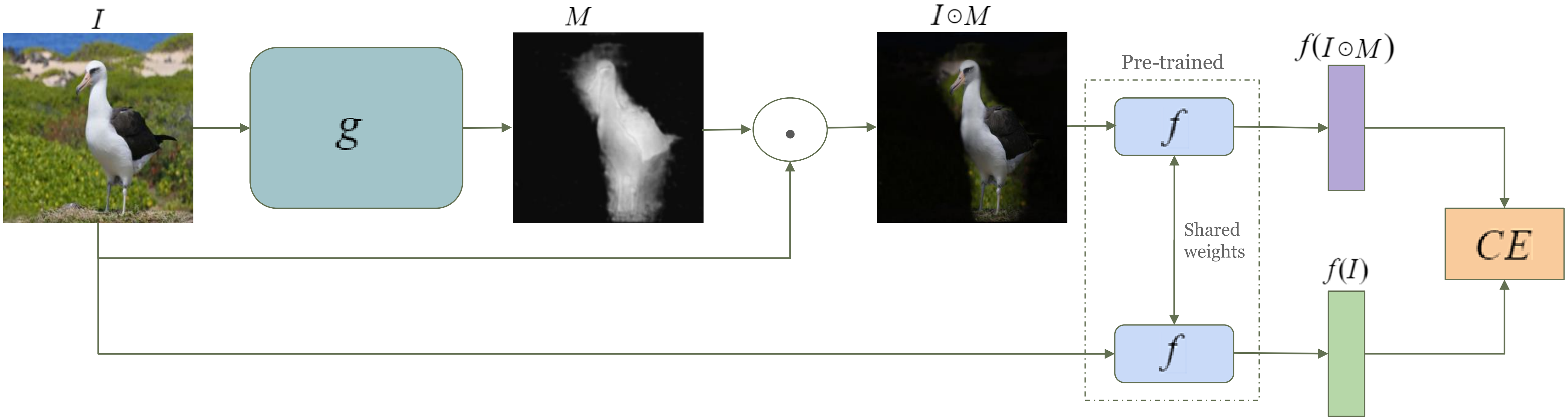
A generic approach that does not assume anything on the AI model



T. Shaharabany, L. Wolf. **Learning a Weight Map for Weakly-Supervised Localization.** ICASSP 23’

Learning a Weight Map

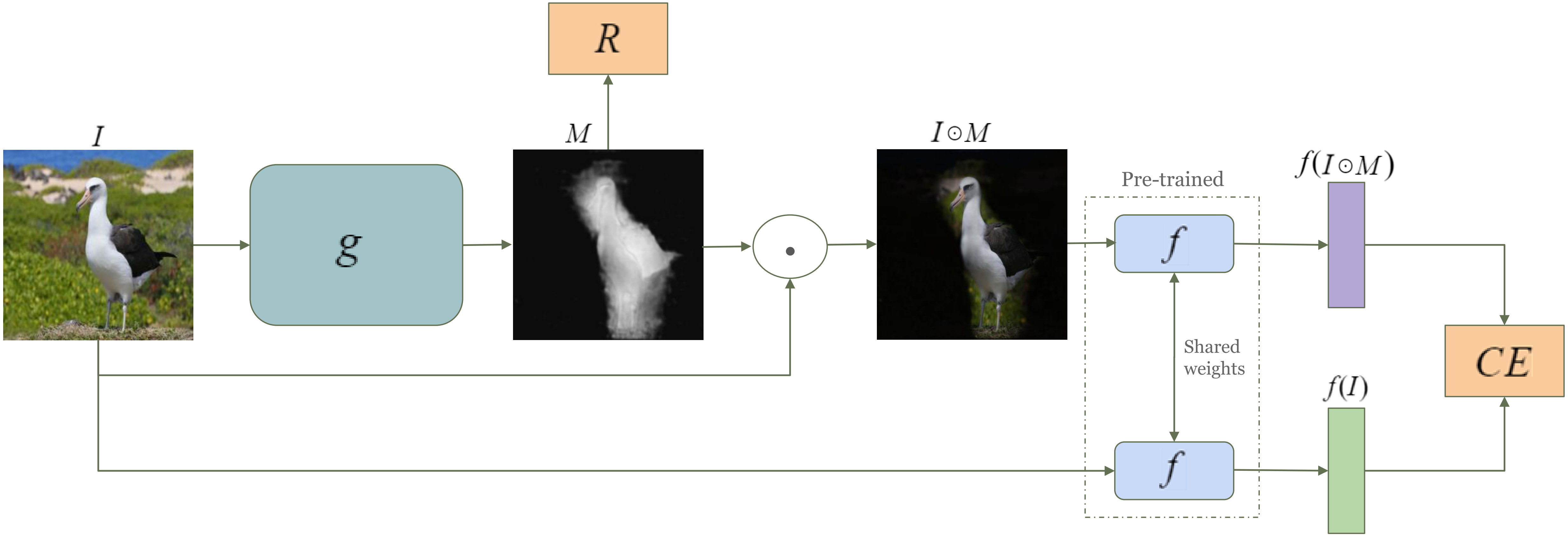
A generic approach that does not assume anything on the AI model



T. Shaharabany, L. Wolf. **Learning a Weight Map for Weakly-Supervised Localization.** ICASSP 23’

Learning a Weight Map

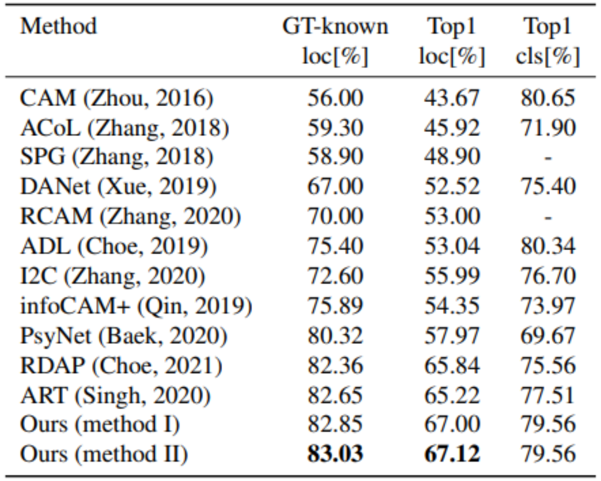
A generic approach that does not assume anything on the AI model



T. Shaharabany, L. Wolf. **Learning a Weight Map for Weakly-Supervised Localization.** ICASSP 23’

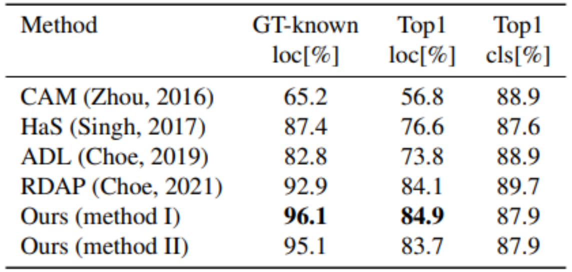
Learning a Weight Map

State of the art in weakly supervised (1) detection and (2) segmentation



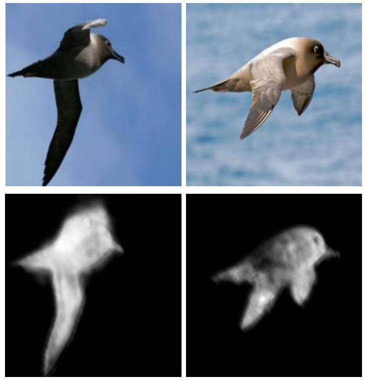
Method	GT-known loc[%]	Top1 loc[%]	Top1 cls[%]
CAM (Zhou, 2016)	56.00	43.67	80.65
ACoL (Zhang, 2018)	59.30	45.92	71.90
SPG (Zhang, 2018)	58.90	48.90	-
DANet (Xue, 2019)	67.00	52.52	75.40
RCAM (Zhang, 2020)	70.00	53.00	-
ADL (Choe, 2019)	75.40	53.04	80.34
I2C (Zhang, 2020)	72.60	55.99	76.70
infoCAM+ (Qin, 2019)	75.89	54.35	73.97
PsyNet (Baek, 2020)	80.32	57.97	69.67
RDAP (Choe, 2021)	82.36	65.84	75.56
ART (Singh, 2020)	82.65	65.22	77.51
Ours (method I)	82.85	67.00	79.56
Ours (method II)	**83.03**	**67.12**	79.56

Table 1. Results on the CUB benchmark

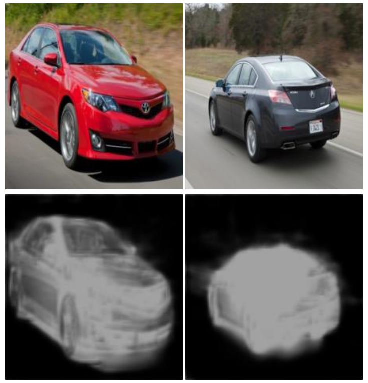


Method	GT-known loc[%]	Top1 loc[%]	Top1 cls[%]
CAM (Zhou, 2016)	65.2	56.8	88.9
HaS (Singh, 2017)	87.4	76.6	87.6
ADL (Choe, 2019)	82.8	73.8	88.9
RDAP (Choe, 2021)	92.9	84.1	89.7
Ours (method I)	**96.1**	**84.9**	87.9
Ours (method II)	95.1	83.7	87.9

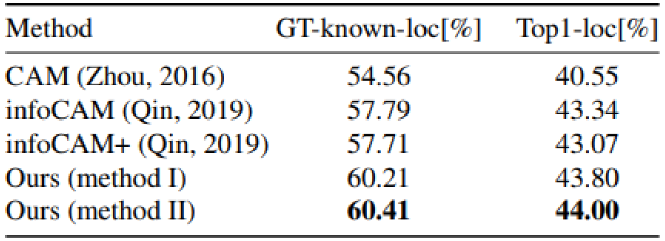
Table 2. Results for the Stanford cars benchmark.



CUB dataset

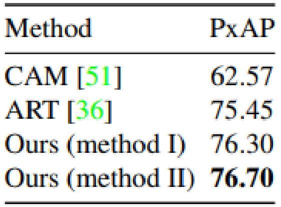


Stanford cars dataset



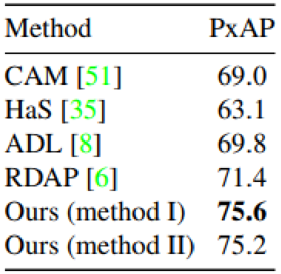
Method	GT-known-loc[%]	Top1-loc[%]
CAM (Zhou, 2016)	54.56	40.55
infoCAM (Qin, 2019)	57.79	43.34
infoCAM+ (Qin, 2019)	57.71	43.07
Ours (method I)	60.21	43.80
Ours (method II)	**60.41**	**44.00**

Table 3. Results for Tiny-imagenet. In all methods, the classifier is a Resnet50.



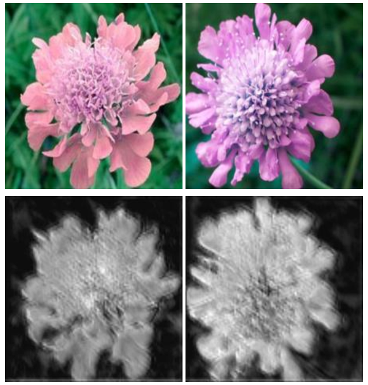
Method	PxAP
CAM [51]	62.57
ART [36]	75.45
Ours (method I)	76.30
Ours (method II)	**76.70**

Table 4. Results for CUB [41] segmentation. The PxAP score aggregates the average precision over multiple thresholds.



Method	PxAP
CAM [51]	69.0
HaS [35]	63.1
ADL [8]	69.8
RDAP [6]	71.4
Ours (method I)	**75.6**
Ours (method II)	75.2

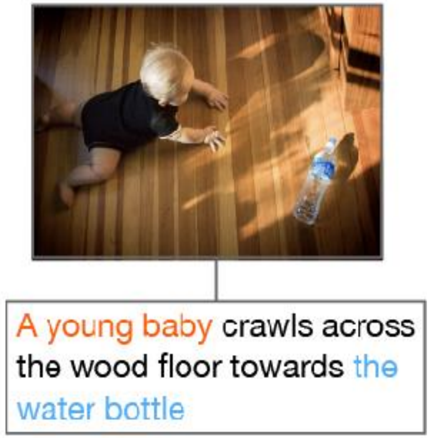
Table 5. Results for oxford flowers segmentation.



Stanford Flowers dataset

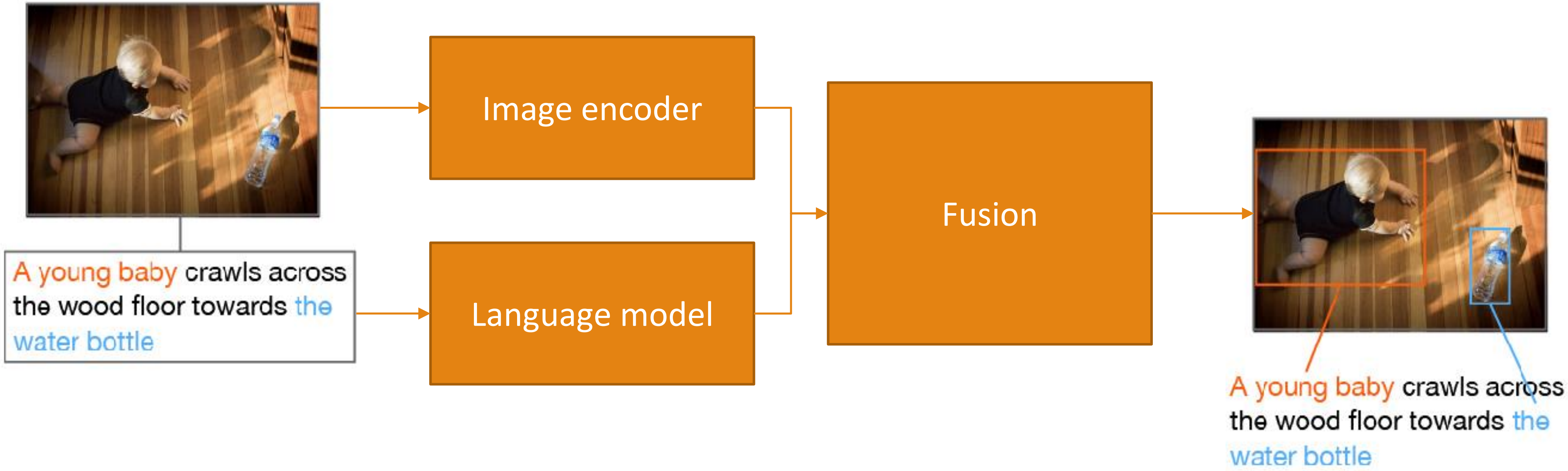
Weakly supervised Phrase-Grounding

Name	Open-world	Weakly supervised	Purely visual
Object Detection	X	X	V
Phrase-Grounding	V	X	X
Weakly supervised localization	X	V	V
Weakly supervised Phrase-Grounding	**V**	**V**	**X**
What is where by looking (WWbL)	V	V	V



A young baby crawls across the wood floor towards the water bottle

Weakly Supervised Phrase Grounding Algorithms



A young baby crawls across the wood floor towards the water bottle

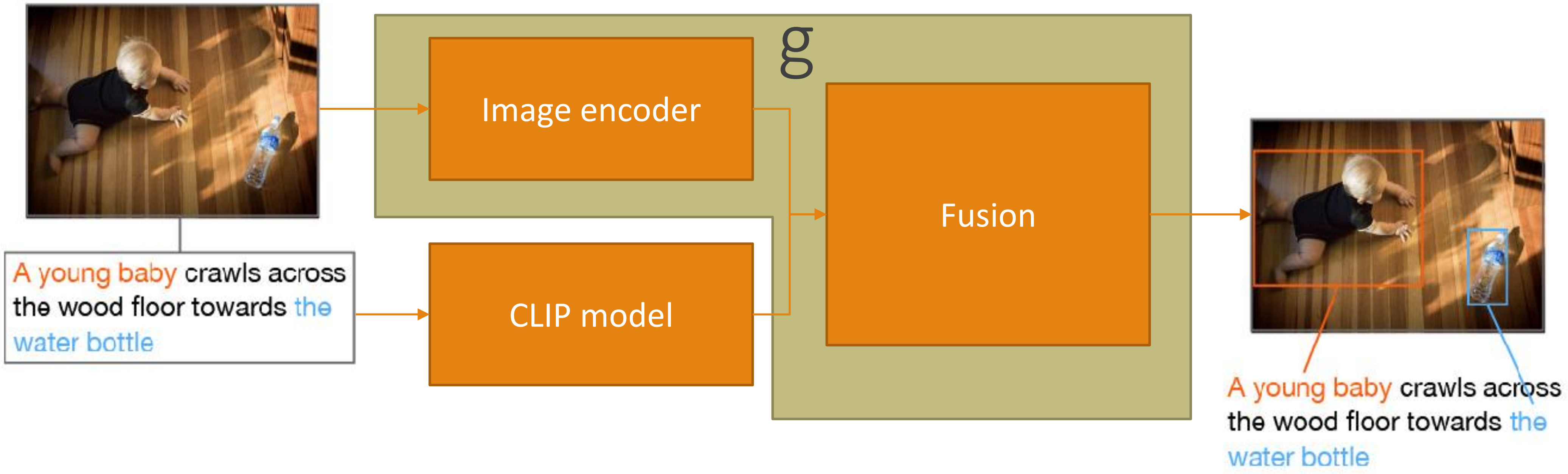
Image encoder

Language model

Fusion

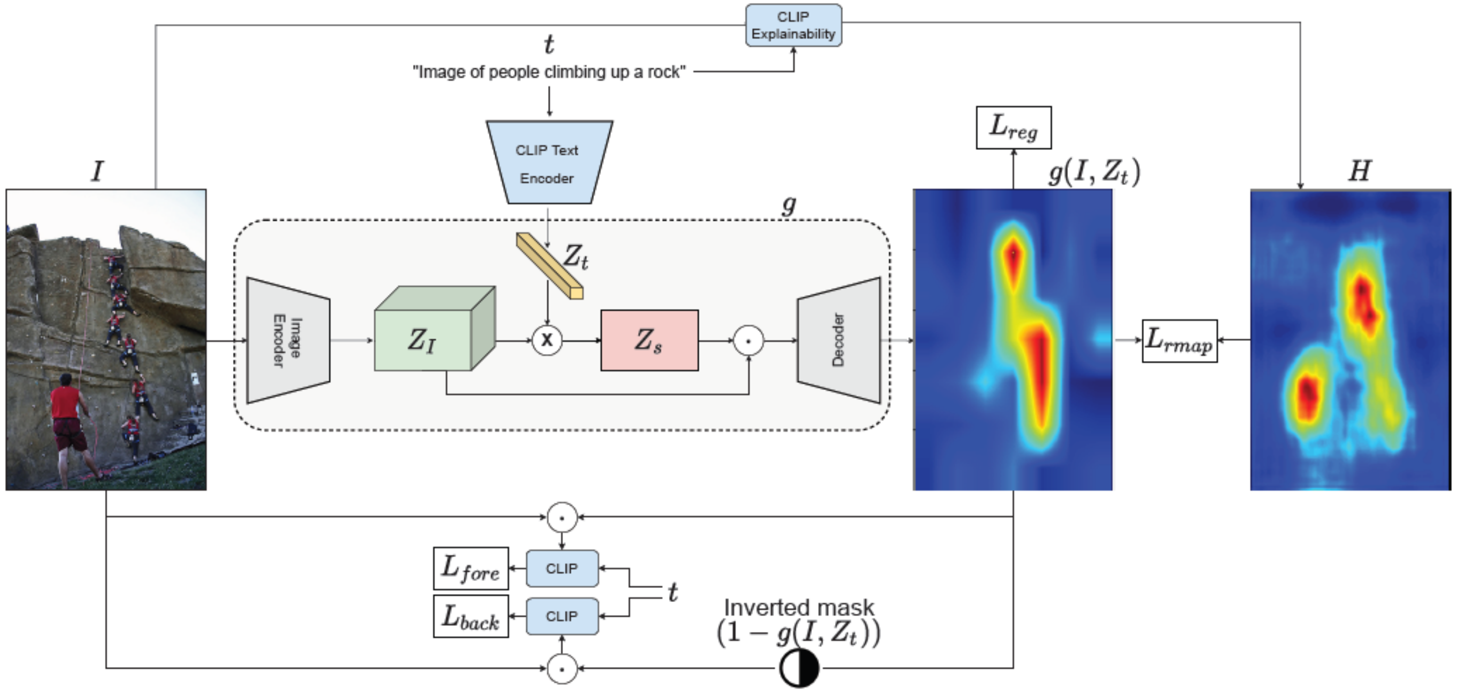
A young baby crawls across the wood floor towards the water bottle

Our Solution - What is Where by Looking (WWbL)


A young baby crawls across the wood floor towards the water bottle
g
Image encoder
CLIP model
Fusion
A young baby crawls across the wood floor towards the water bottle

T. Shaharabany, Y. Tewel, L. Wolf. **What is Where by Looking (WWbL) – Weakly-Supervised Open-World Phrase-Grounding without Text Inputs.** NeurIPS'22

Architecture - What is Where by Looking (WWbL)



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

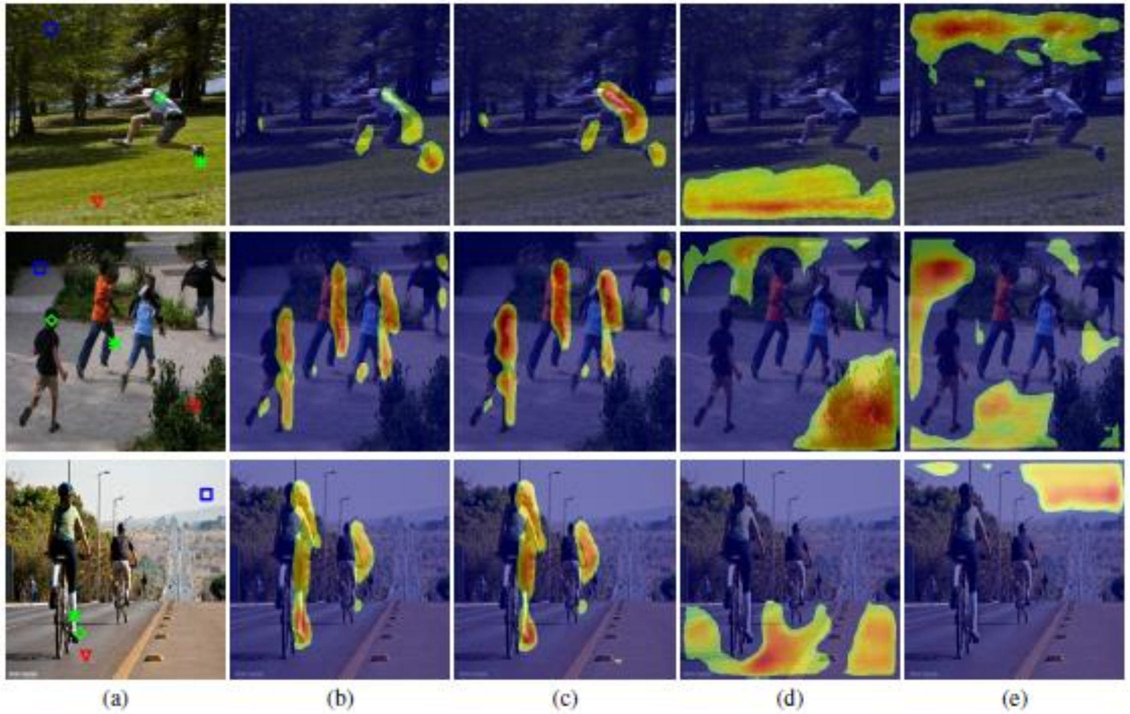
$$L_{back}(I,t) = CLIP((1-g(I,Z^T)) \odot I,t).$$

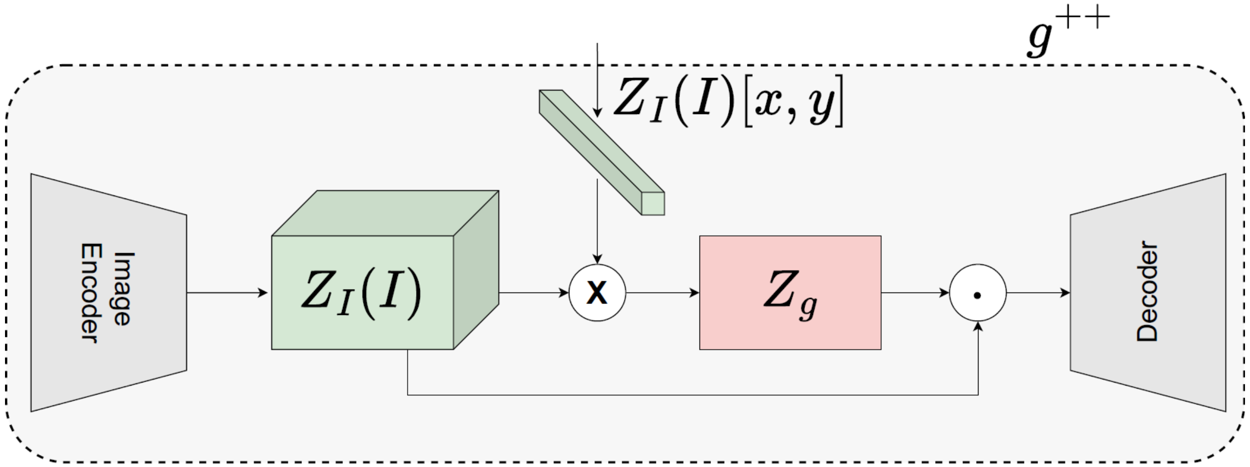
$$L_{rmap}(I,H) = \|H - g(I,Z^T)\|^2$$

$$L_{reg}(I,g)) = \|g(I,Z^T)\|$$

T. Shaharabany, Y. Tewel, L. Wolf. **What is Where by Looking (WWbL) – Weakly-Supervised Open-World Phrase-Grounding without Text Inputs.** NeurIPS’22

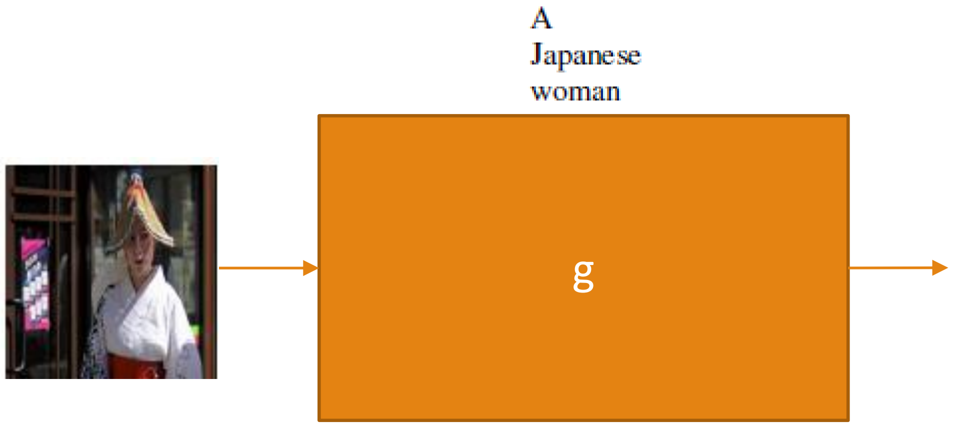
Similarity Maps for Phrase Grounding





T. Shaharabany, L. Wolf. **Similarity Maps for Self-Training Weakly-Supervised Phrase Grounding.** CVPR'23

Method – Maps Selection

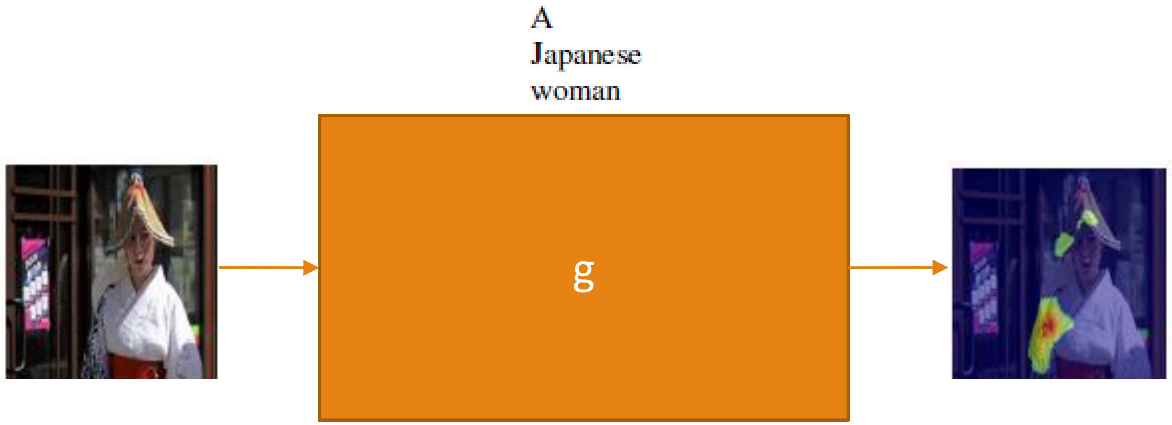


A
Japanese
woman

g

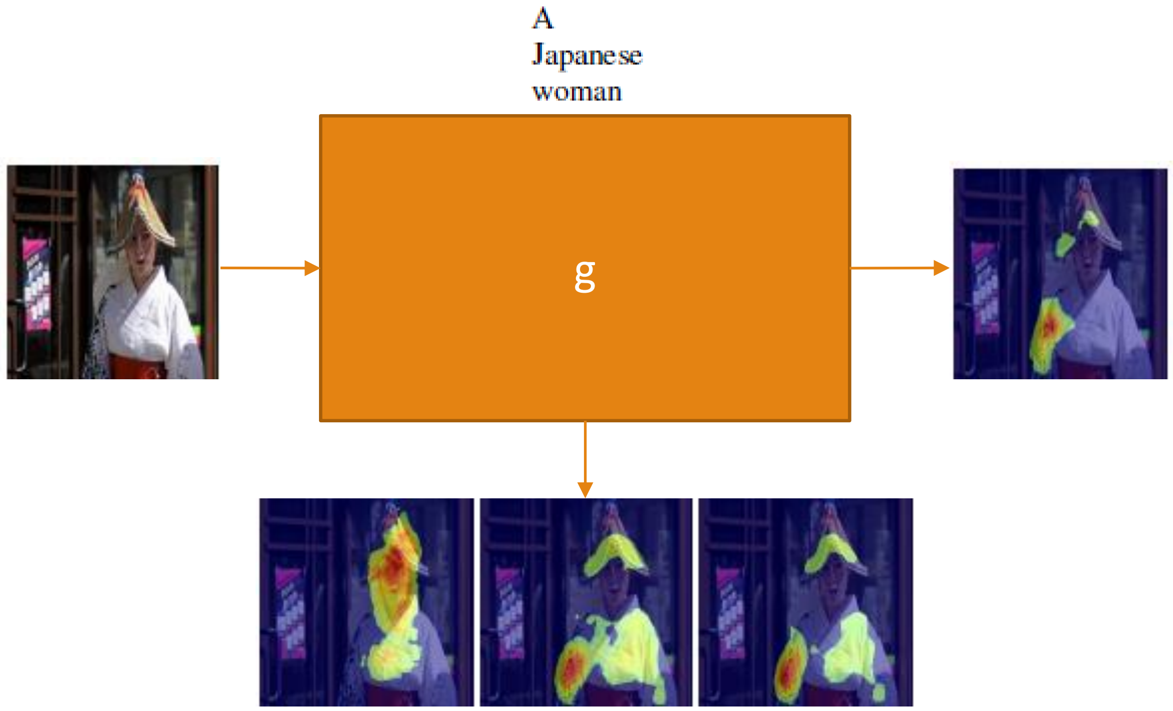
T. Shaharabany, L. Wolf. **Similarity Maps for Self-Training Weakly-Supervised Phrase Grounding.** CVPR’23

Method – Maps Selection



T. Shaharabany, L. Wolf. **Similarity Maps for Self-Training Weakly-Supervised Phrase Grounding.** CVPR’23

Method – Maps Selection



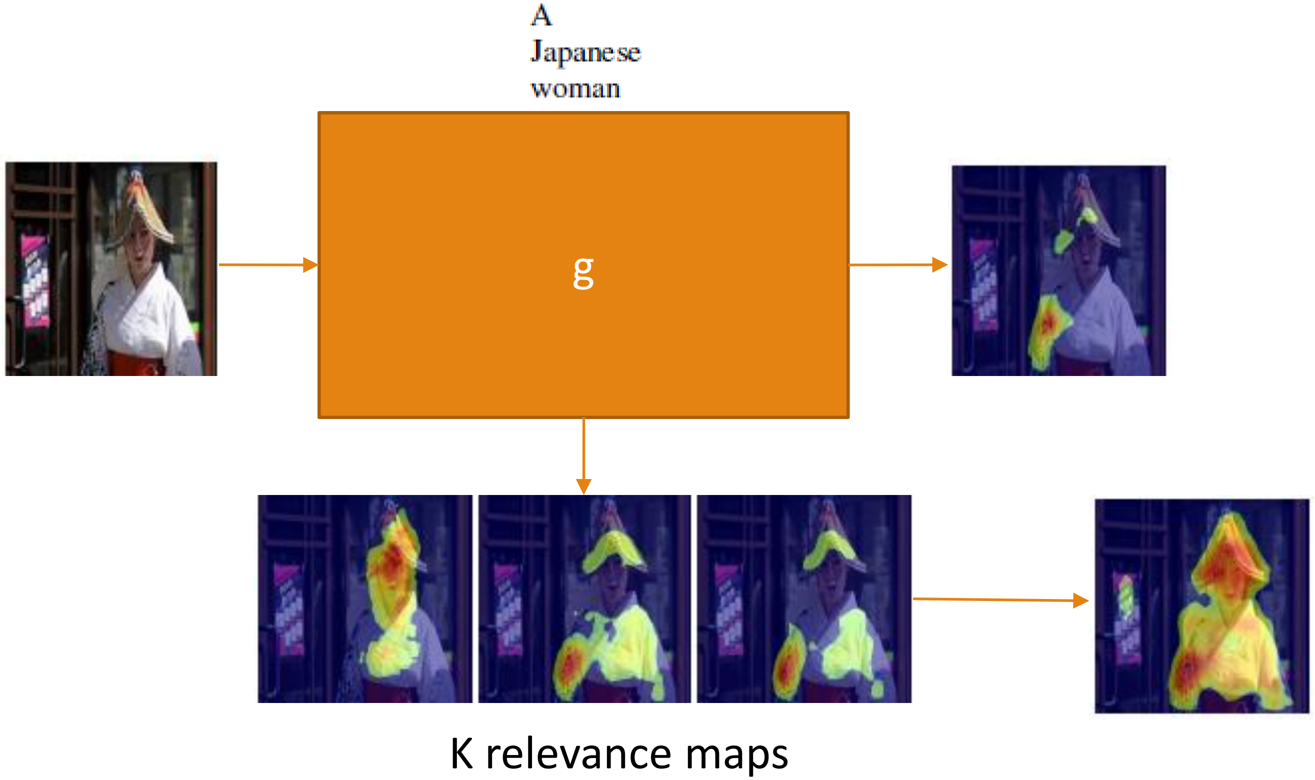
A Japanese woman

g

K relevance maps

T. Shaharabany, L. Wolf. **Similarity Maps for Self-Training Weakly-Supervised Phrase Grounding.** CVPR'23

Method – Maps Selection



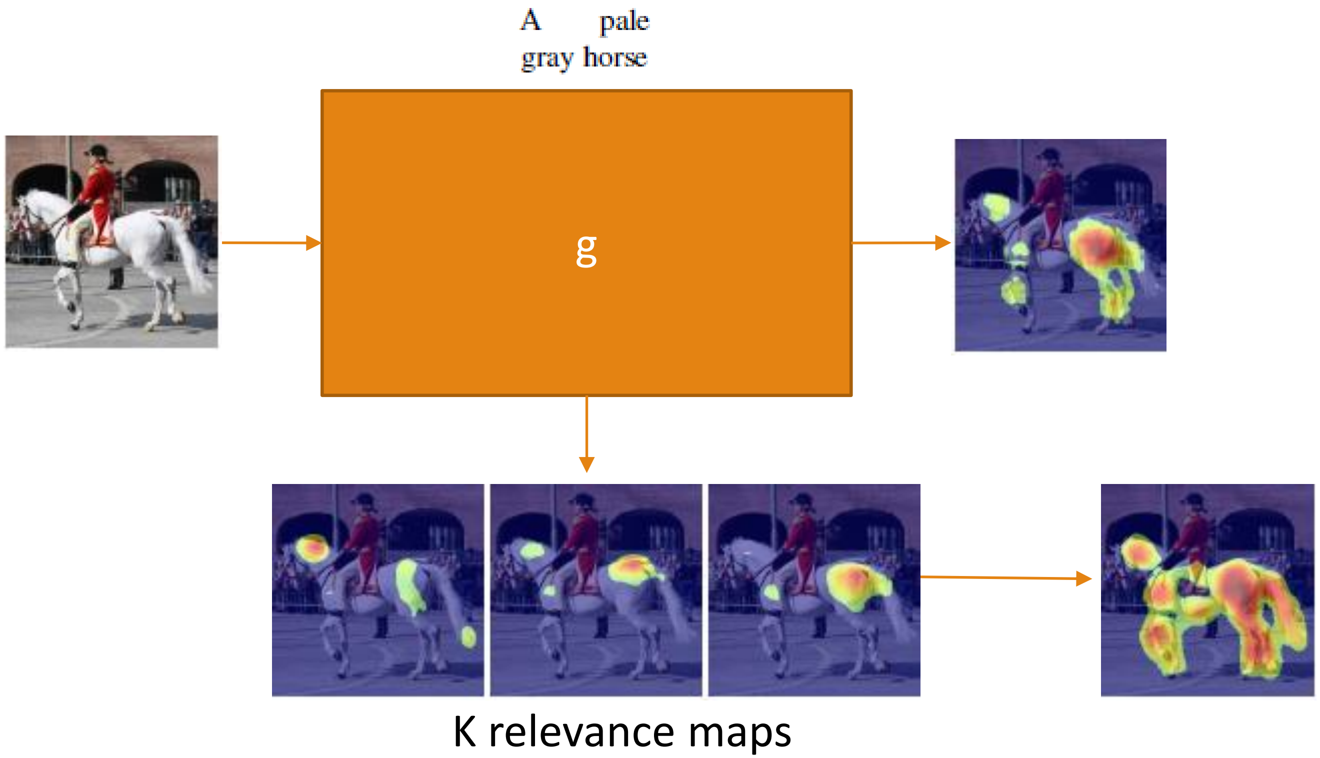
A Japanese woman

g

K relevance maps

T. Shaharabany, L. Wolf. **Similarity Maps for Self-Training Weakly-Supervised Phrase Grounding.** CVPR'23

Method – Maps Selection



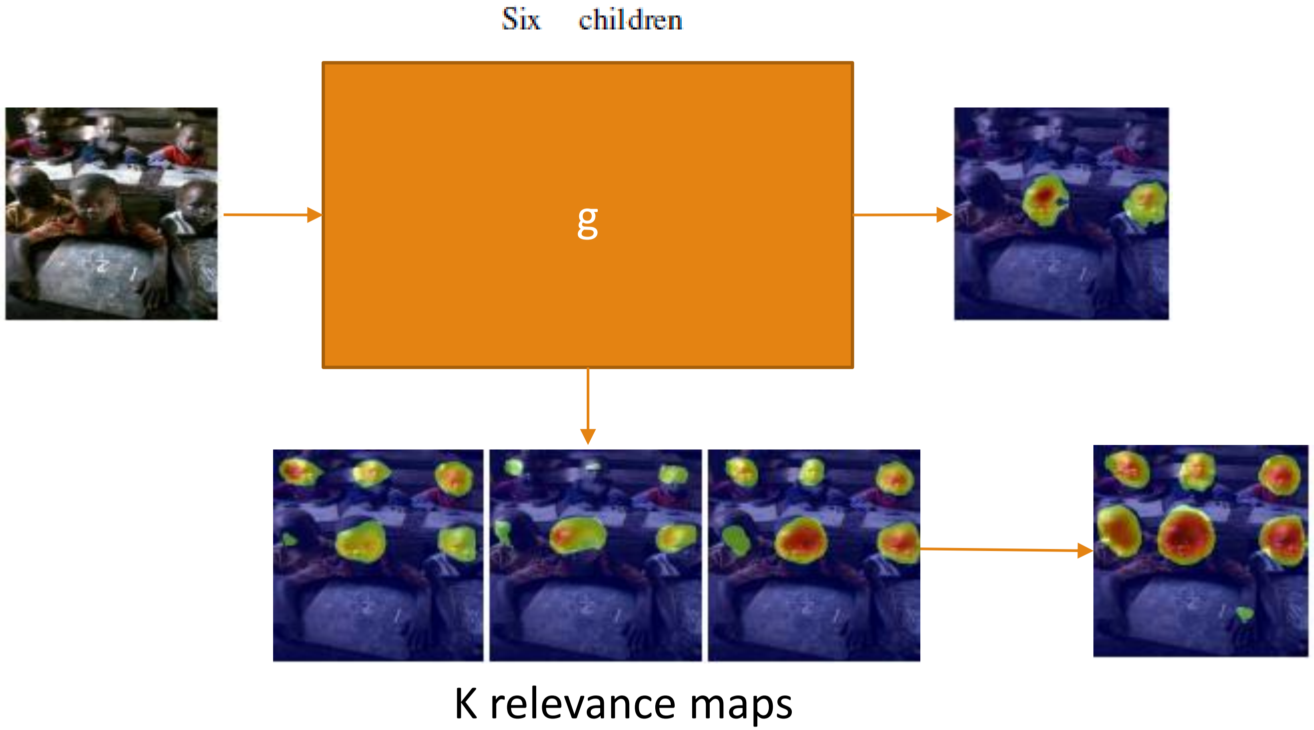
A pale gray horse

g

K relevance maps

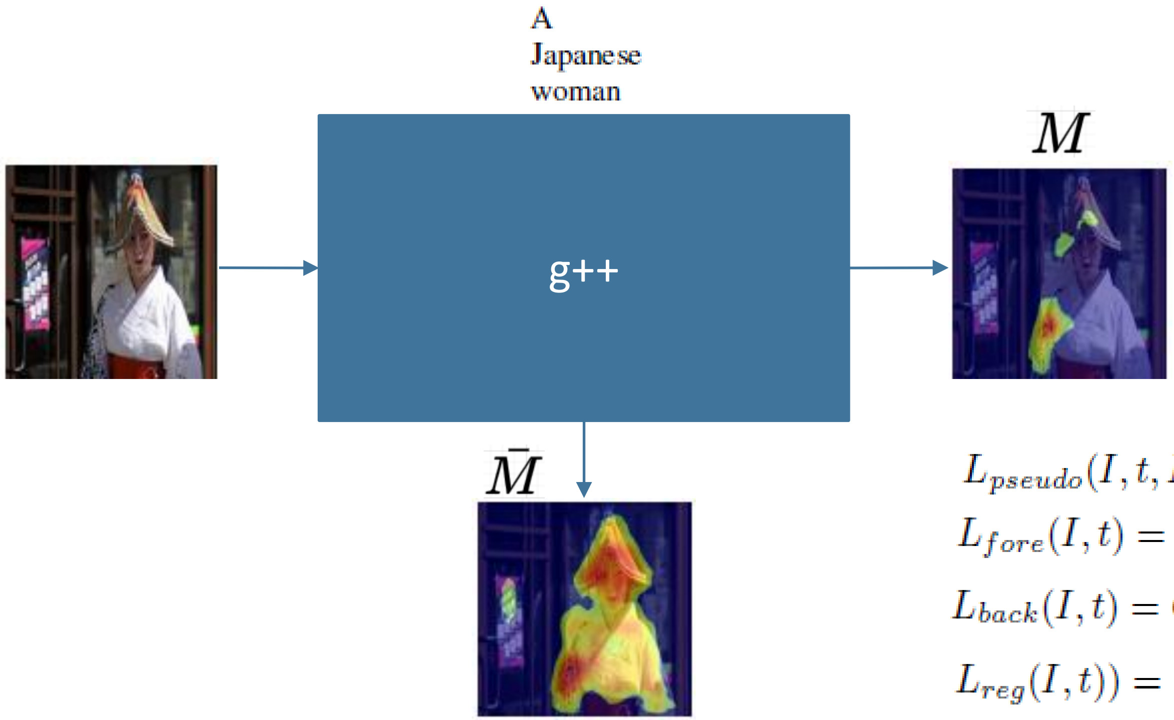
T. Shaharabany, L. Wolf. **Similarity Maps for Self-Training Weakly-Supervised Phrase Grounding.** CVPR’23

Method – Maps Selection



T. Shaharabany, L. Wolf. **Similarity Maps for Self-Training Weakly-Supervised Phrase Grounding.** CVPR'23

Method – Fine-tune



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

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

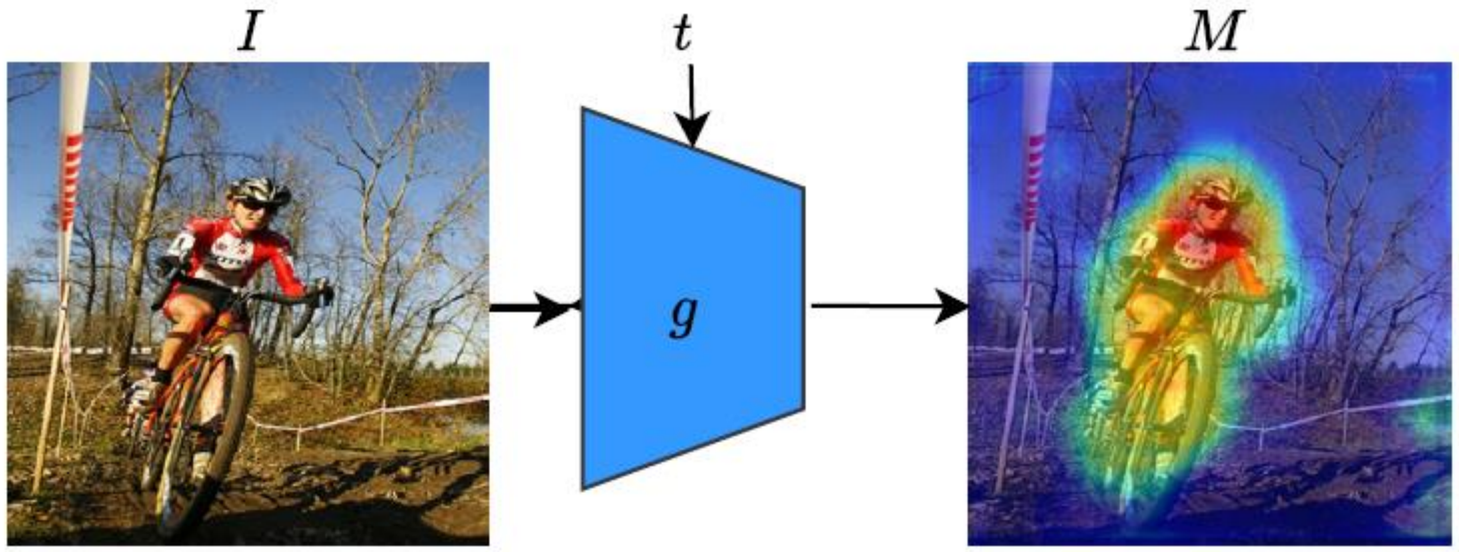
$$L_{back}(I,t) = CLIP((1 - g^{++}(I, Z_t(t))) \odot I, t).$$

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

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

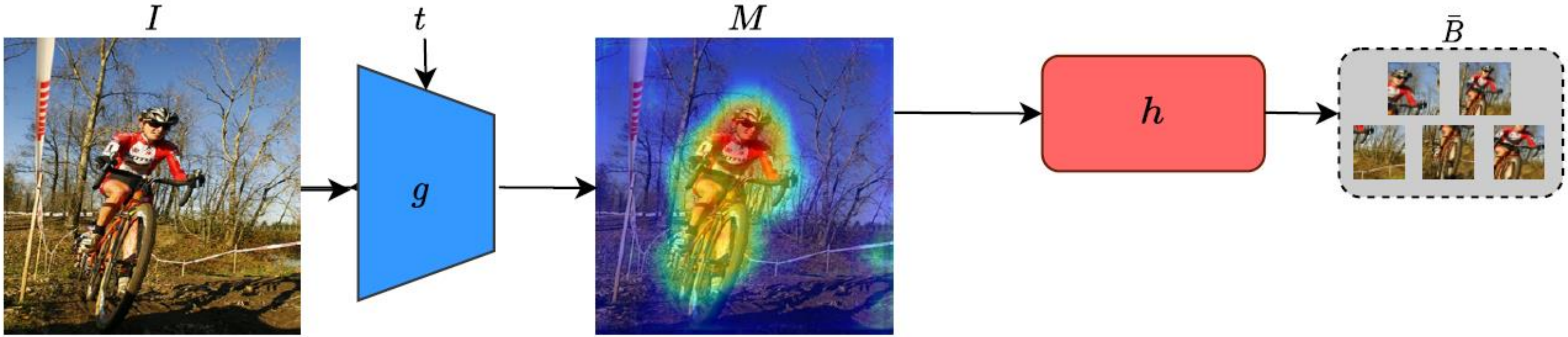
T. Shaharabany, L. Wolf. **Similarity Maps for Self-Training Weakly-Supervised Phrase Grounding.** CVPR’23

Box-based Refinement for Phrase Grounding



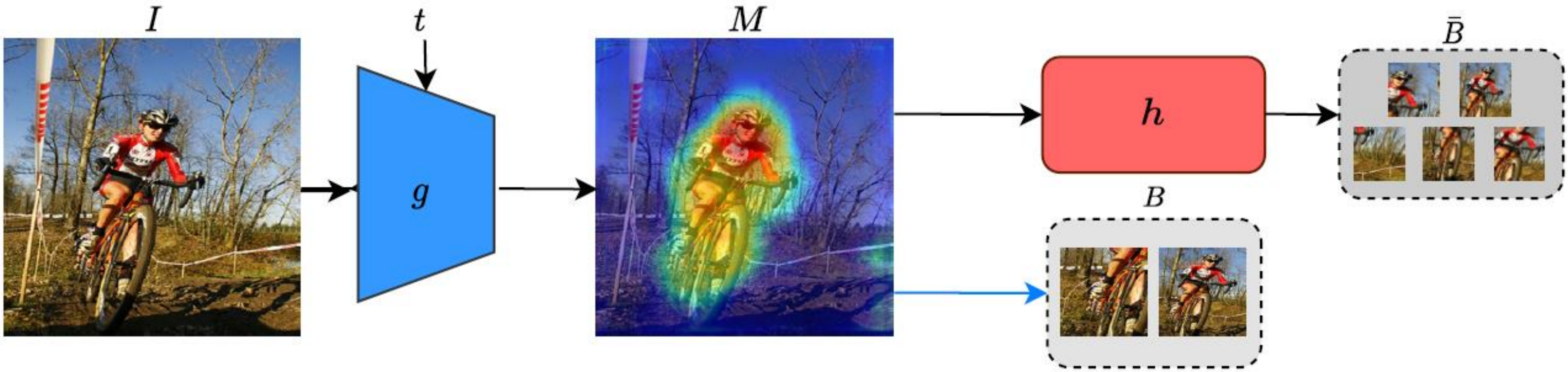
E. Gomel, T. Shaharabany, L. Wolf. **Box-based Refinement for Weakly Supervised and Unsupervised Localization Tasks.** In submission

Box-based Refinement for Phrase Grounding



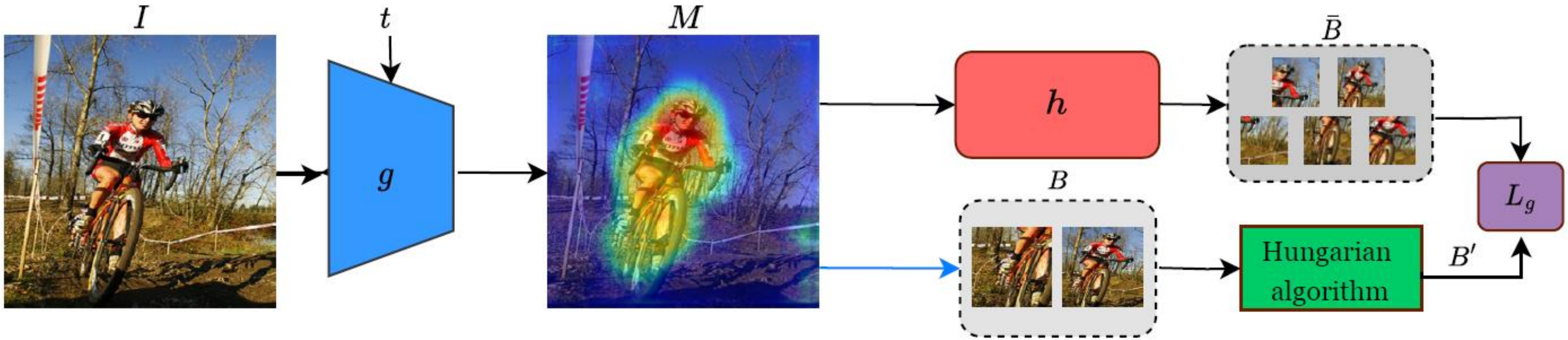
E. Gomel, T. Shaharabany, L. Wolf. **Box-based Refinement for Weakly Supervised and Unsupervised Localization Tasks.** In submission

Box-based Refinement for Phrase Grounding



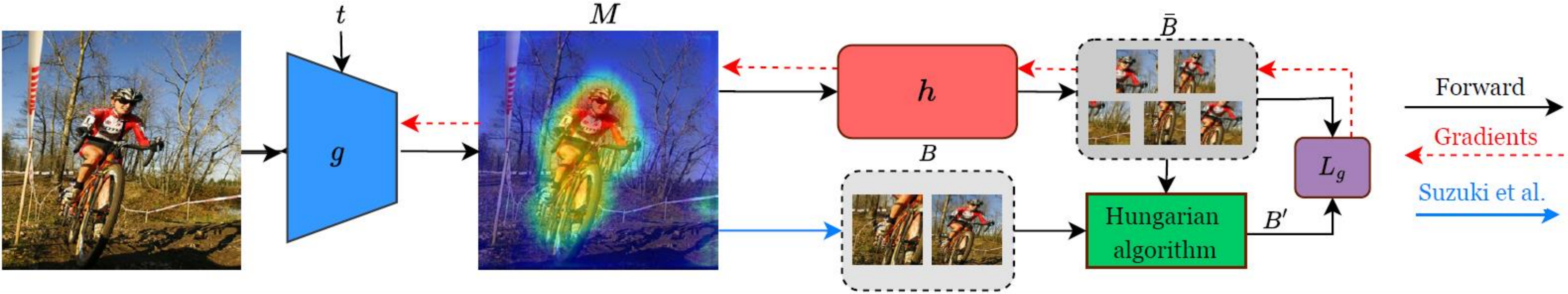
E. Gomel, T. Shaharabany, L. Wolf. **Box-based Refinement for Weakly Supervised and Unsupervised Localization Tasks.** In submission

Box-based Refinement for Phrase Grounding



E. Gomel, T. Shaharabany, L. Wolf. **Box-based Refinement for Weakly Supervised and Unsupervised Localization Tasks.** In submission

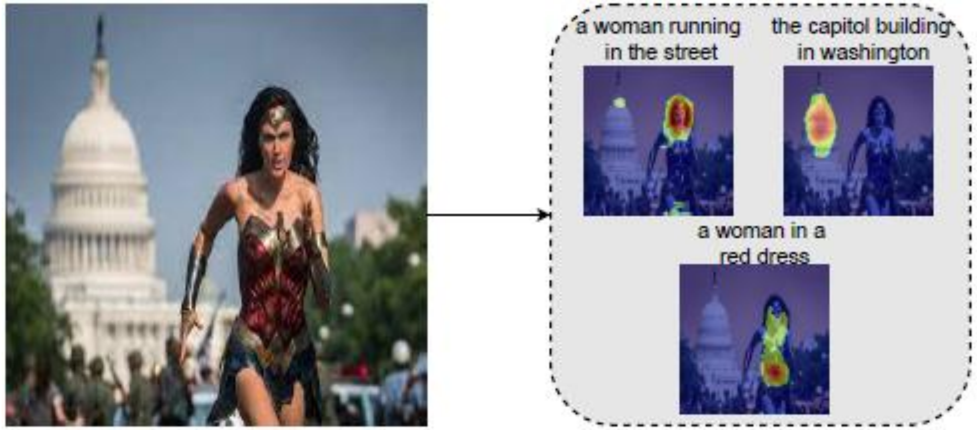
Box-based Refinement for Phrase Grounding



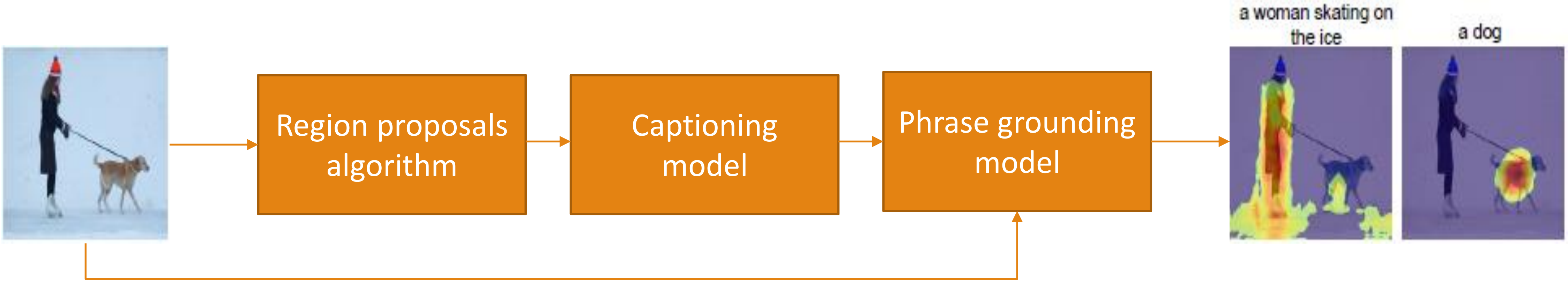
E. Gomel, T. Shaharabany, L. Wolf. **Box-based Refinement for Weakly Supervised and Unsupervised Localization Tasks.** In submission

What is where by looking (WWbL)

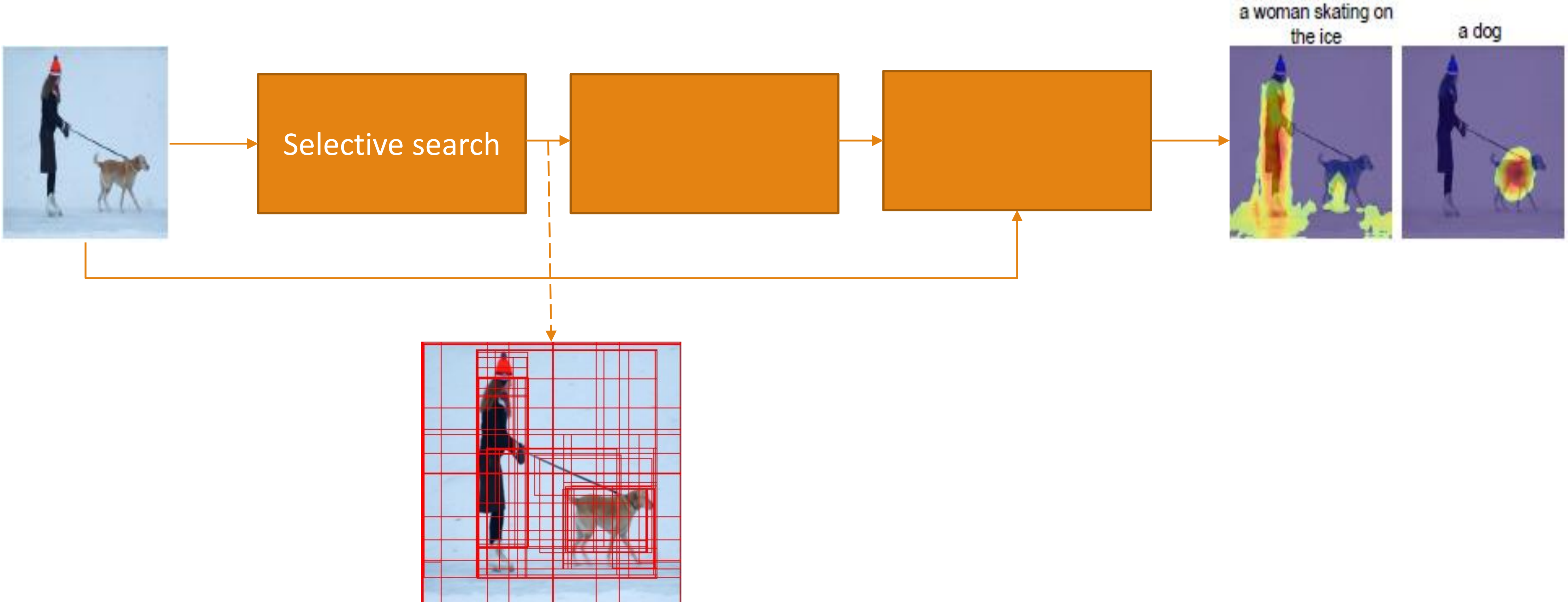
Name	Open-world	Weakly supervised	Purely visual
Object Detection	X	X	V
Phrase-Grounding	V	X	X
Weakly supervised localization	X	V	V
Weakly supervised Phrase-Grounding	V	V	X
What is where by looking (WWbL)	**V**	**V**	**V**


a woman running in the street
the capitol building in washington
a woman in a red dress

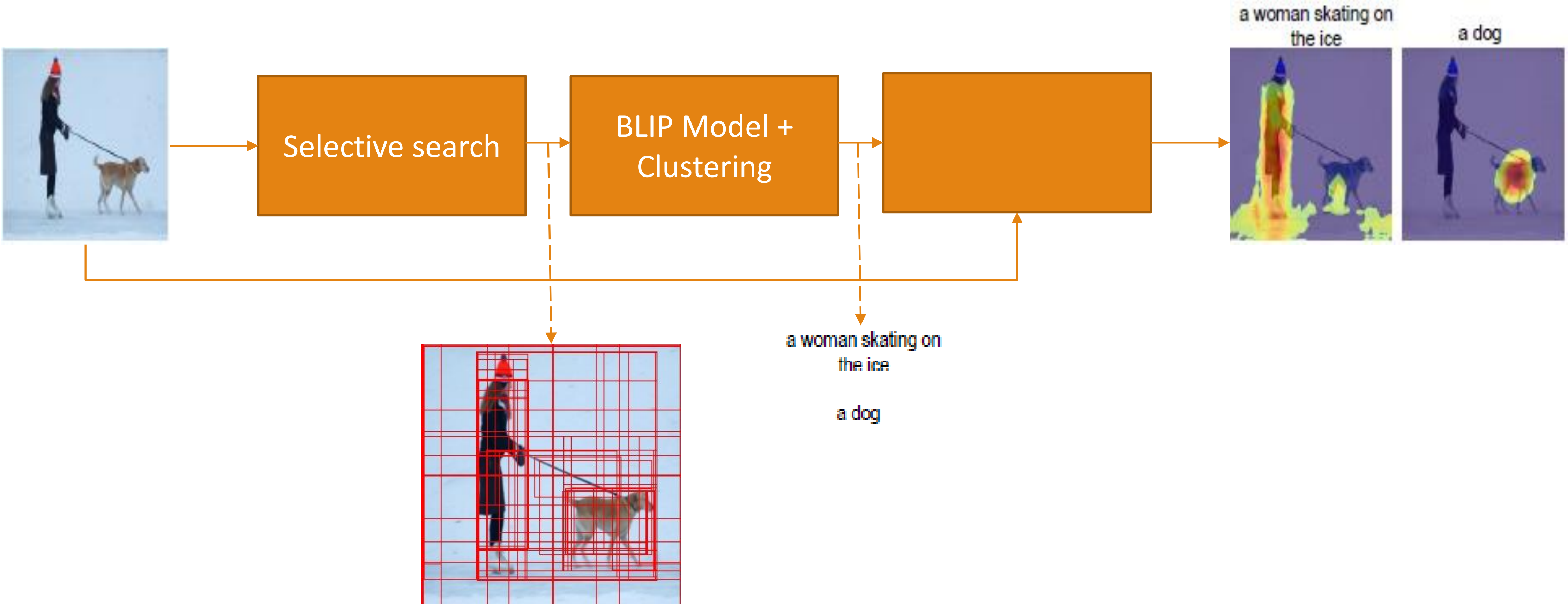
WWbL Algorithms



Proposed Algorithm



Proposed Algorithm



Selective search

BLIP Model + Clustering

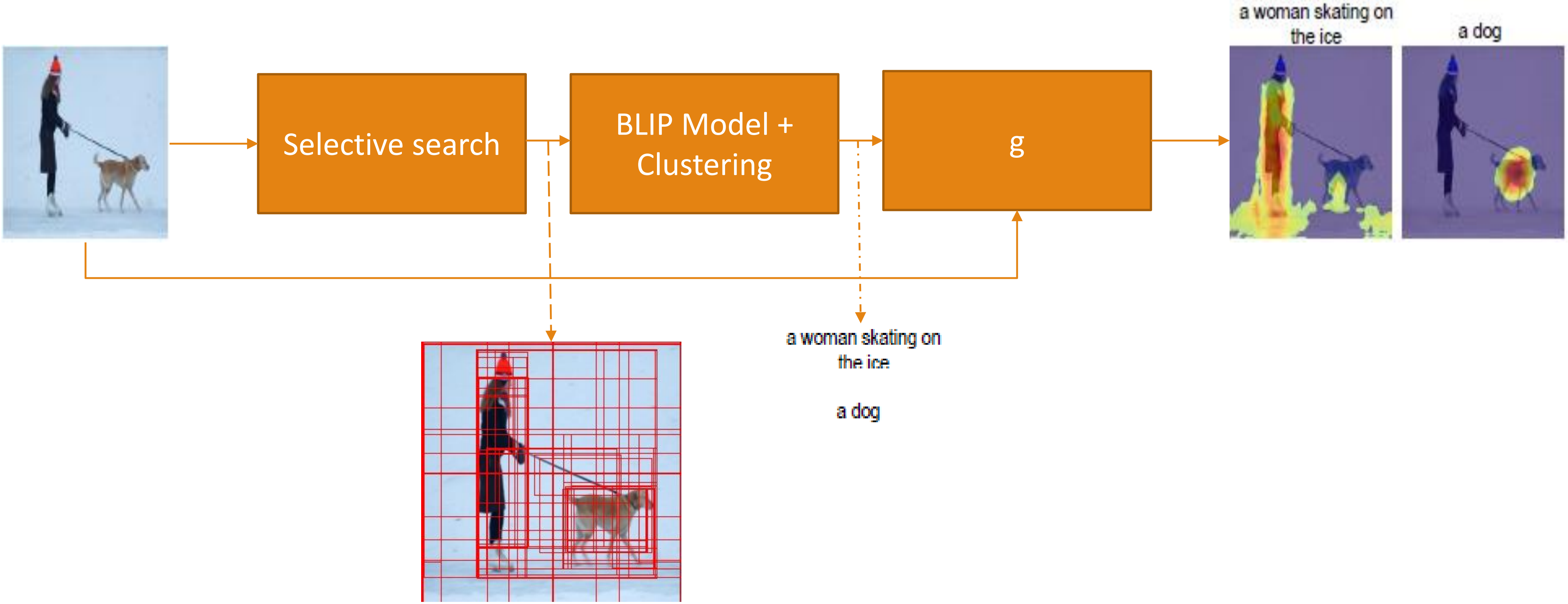
a woman skating on the ice

a dog

a woman skating on the ice

a dog

Proposed Algorithm



Selective search

BLIP Model + Clustering

g

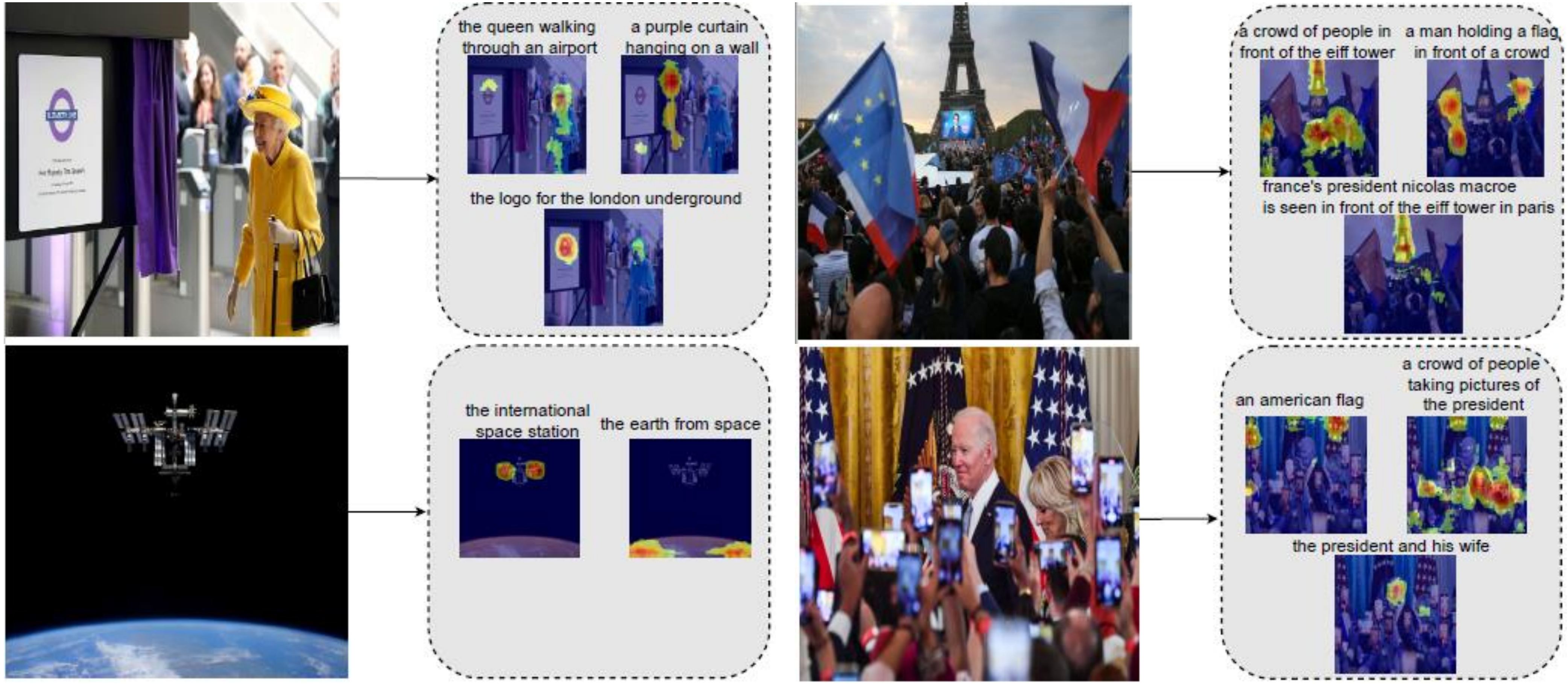
a woman skating on the ice

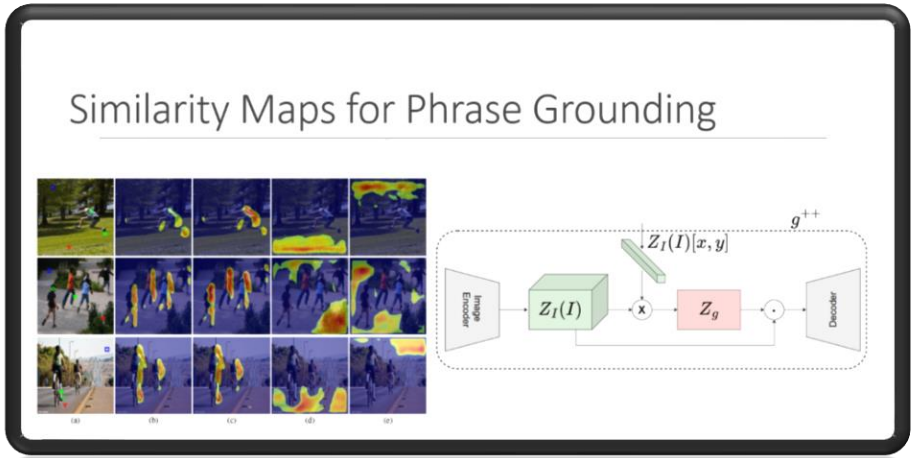
a dog

a woman skating on the ice

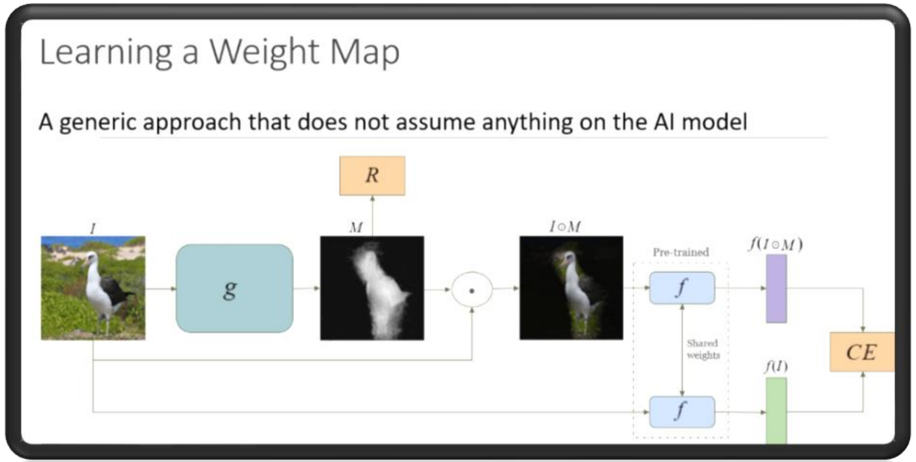
a dog

Visualization – What is Where by Looking


the queen walking through an airport
a purple curtain hanging on a wall
the logo for the london underground
a crowd of people in front of the eiff tower
a man holding a flag in front of a crowd
france's president nicolas macroe is seen in front of the eiff tower in paris
the international space station
the earth from space
an american flag
a crowd of people taking pictures of the president
the president and his wife

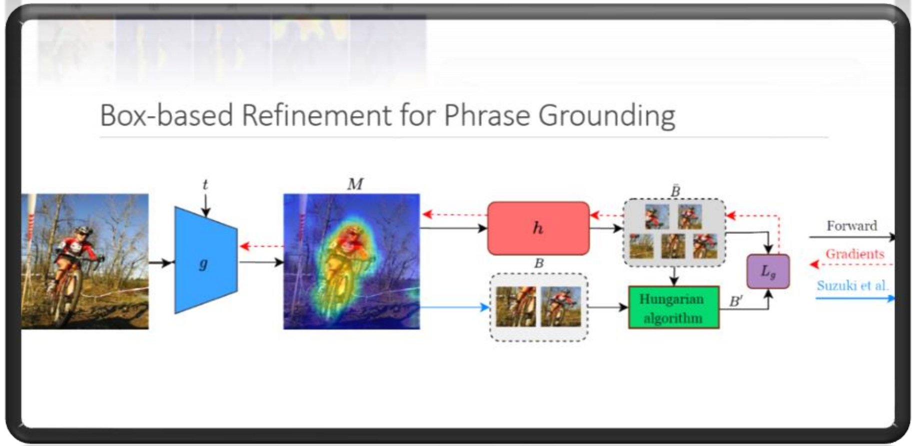


Similarity Maps for Phrase Grounding

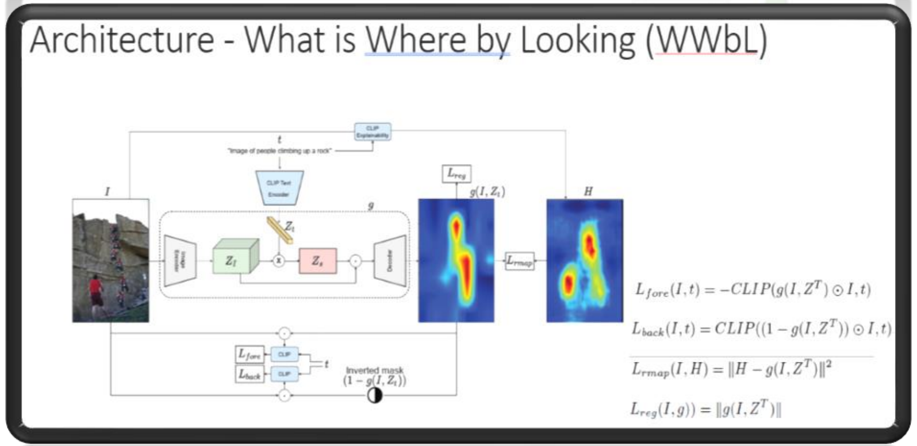


Learning a Weight Map

A generic approach that does not assume anything on the AI model



Box-based Refinement for Phrase Grounding



Architecture - What is Where by Looking (WWbL)

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

$$L_{back}(I,t) = CLIP((1 - g(I,Z^T)) \odot I, t)$$

$$L_{rmap}(I,H) = \|H - g(I,Z^T)\|^2$$

$$L_{reg}(I,g)) = \|g(I,Z^T)\|$$

Thank you