JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Learning Multi-Modal Class-Specific Tokens for Weakly Supervised Dense Object Localization

Lian Xu[1], Wanli Ouyang[2], Mohammed Bennamoun[1], Farid Boussaid[1], Dan Xu[3]

Poster: THU-AM-296

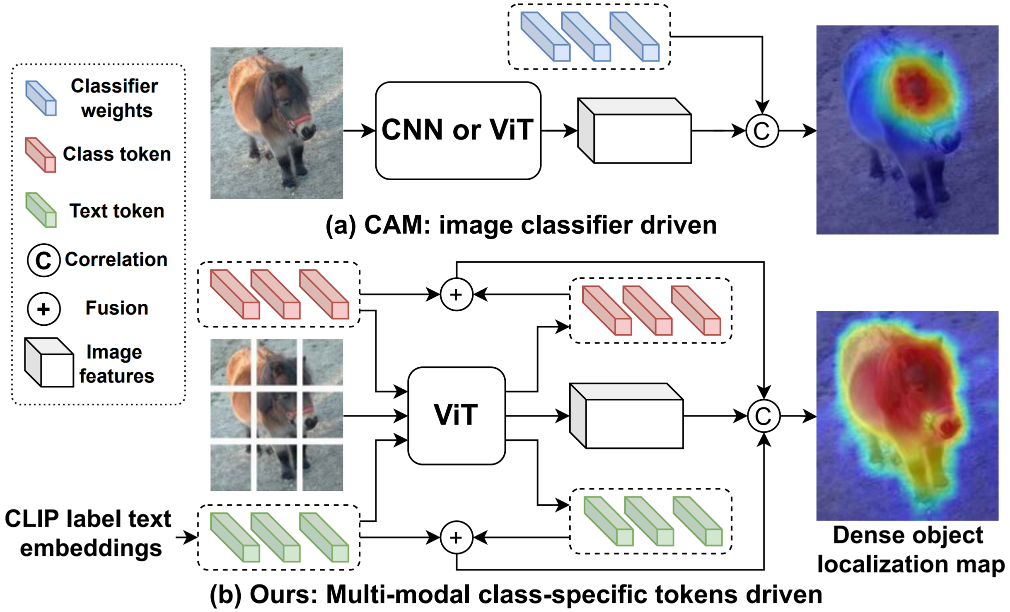
[1]The University of Western Australia, [2]Shanghai AI Laboratory
[3]Hong Kong University of Science and Technology

THE UNIVERSITY OF WESTERN AUSTRALIA

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

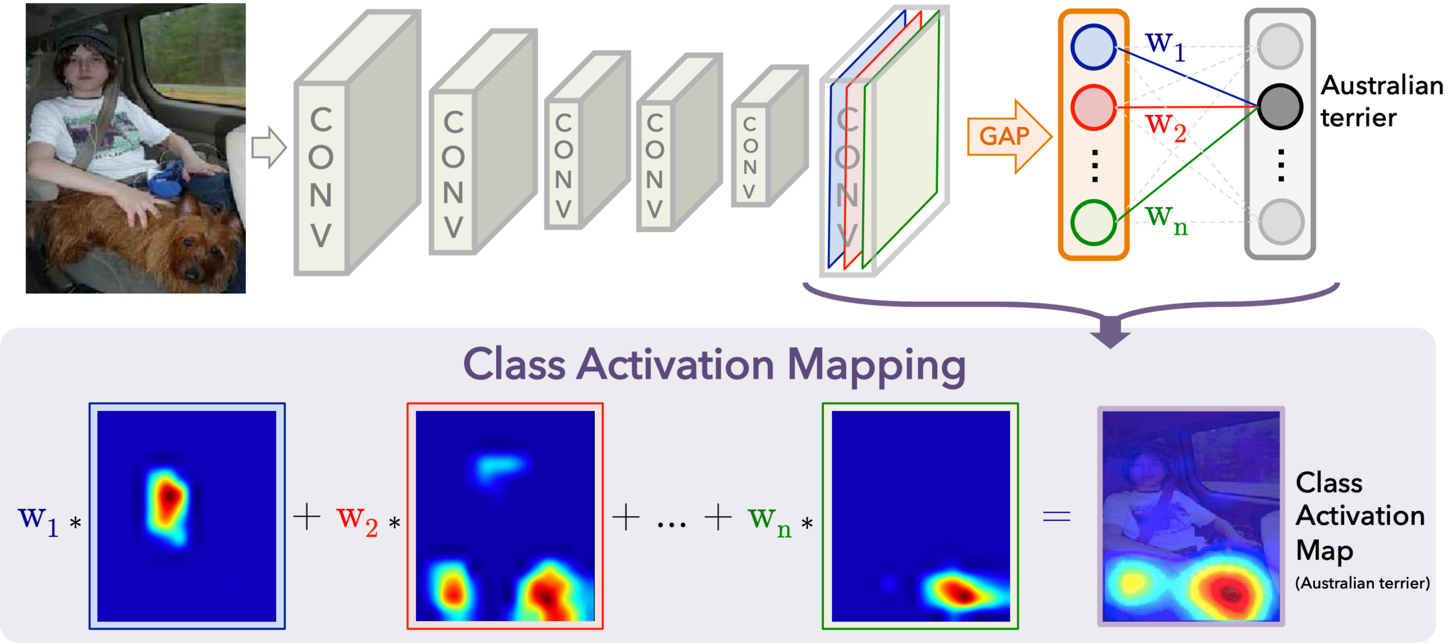
Overview



We focus on learning class representations that can well correlate pixel features for accurate dense object localization.

- We propose to explicitly construct multi-modal class representations in a unified transformer framework.
- We propose to learn class-specific visual and textual tokens by leveraging the pre-trained CLIP model
- We propose to enhance the multi-modal class-specific tokens by incorporating sample-specific context
- The proposed WSDOL results lead to SoTA WSSS results on PASCAL VOC and MS COCO.

Weakly Supervised Dense Object Localization



CAM mechanism:

$$\boldsymbol{A}_i^c = \sum_{k=1}^{K} \boldsymbol{w}_c^k \boldsymbol{F}_i^k,$$

i.e., the correlation between class-specific weights of the image classifier ($\boldsymbol{w}_c$) and pixel-level features ($\boldsymbol{F}_i$)

Limitations:

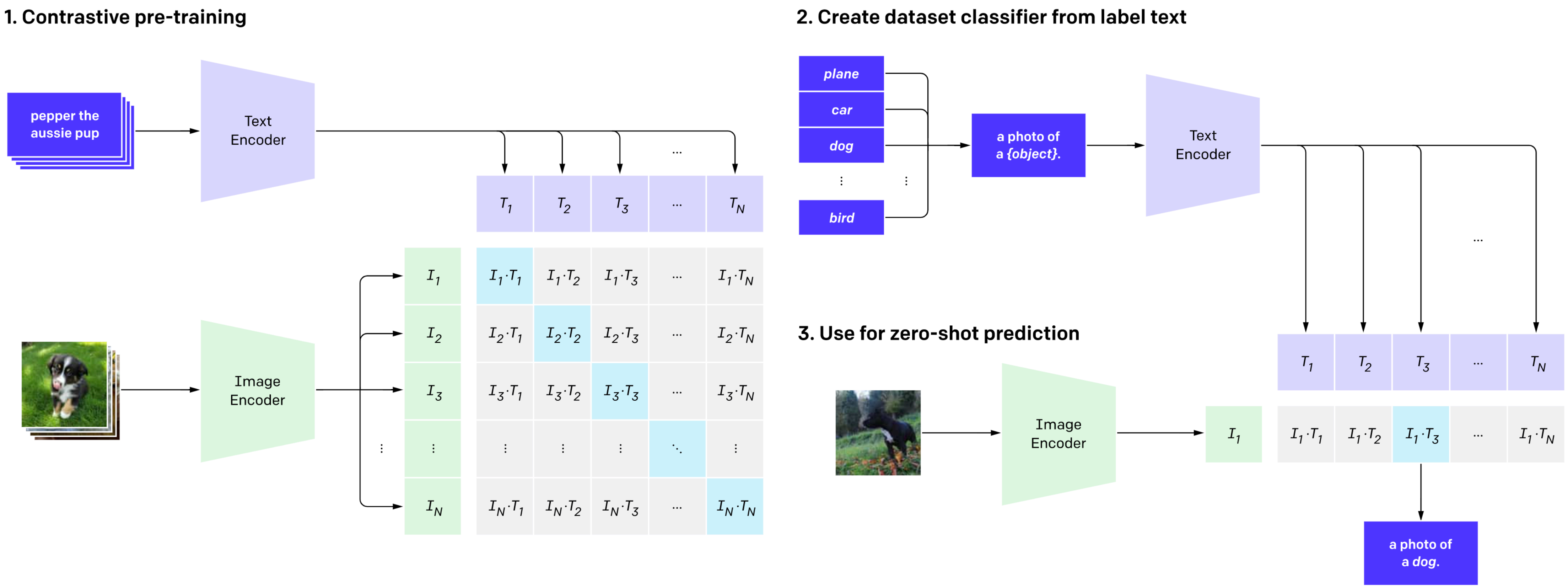
The class-specific weights are class representations, which are

- image classification representations, with a limited ability to address intra-class variations;
- global dataset-level representations, not adaptive to capture sample-specific features;

resulting in inaccurate class-to-pixel correlation.

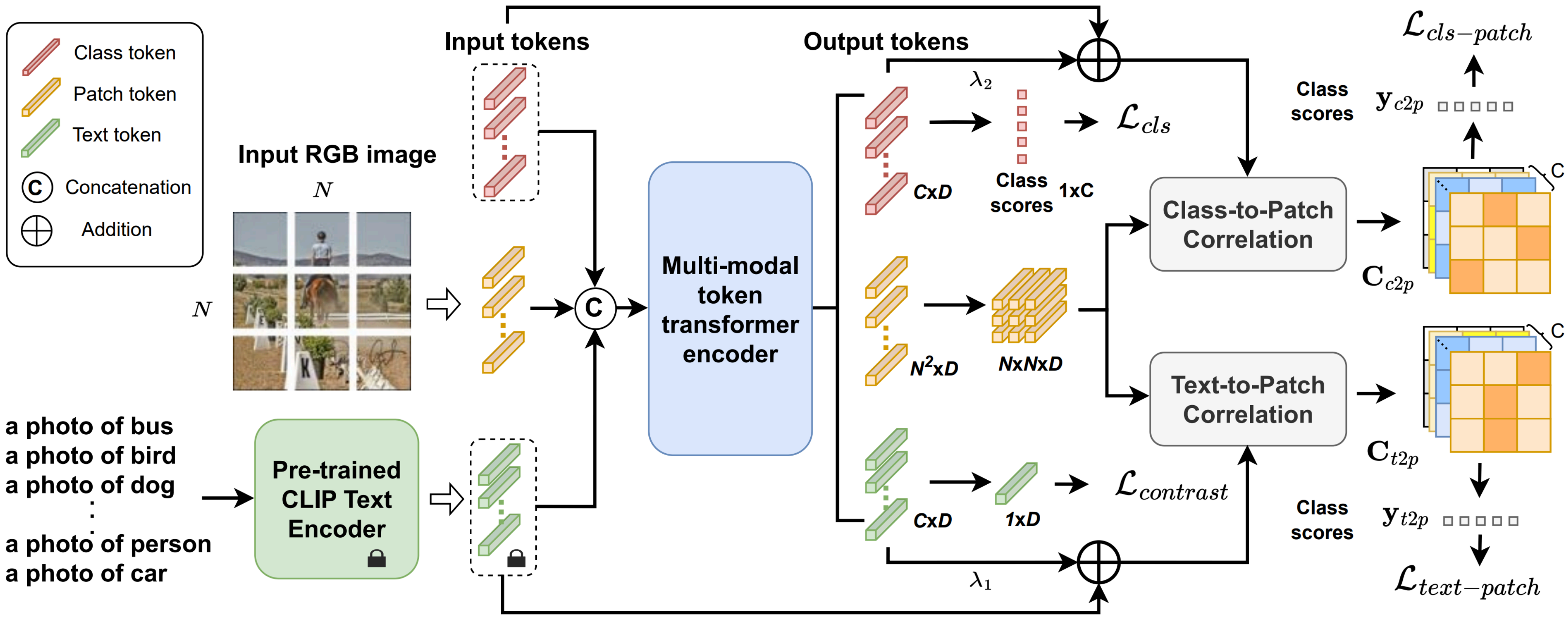
Goal: To learn more discriminative and sample-adaptive class representations for dense object localization.

Contrastive Language-Image Pretraining (CLIP)



It provides a novel way to learn visual concepts through natural language supervisions.

The proposed framework



Multi-modal class-specific token learning

- Class-specific **textual** tokens:

$$T_{txt} = T_{txt}^{in} + \lambda_1 \cdot T_{txt}^{out}$$

T_{txt}^{in}: *global* class-specific textual tokens, initialized by the pre-trained CLIP label text embeddings.

T_{txt}^{out}: *local* class-specific textual tokens, refined by *sample-specific visual context*.

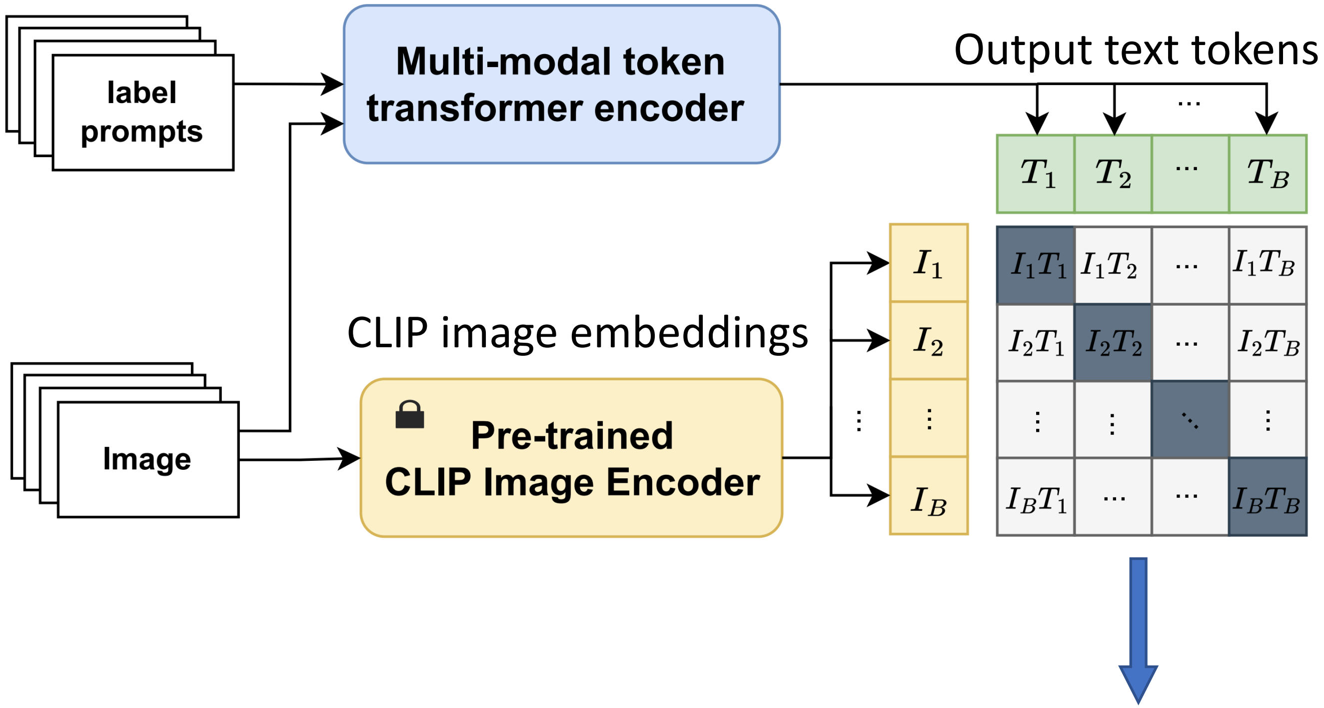
- Class-specific **visual** tokens:

$$T_{cls} = T_{cls}^{in} + \lambda_2 \cdot T_{cls}^{out}$$

T_{cls}^{in} : *global* class-specific visual tokens, initialized by the pre-trained DINO class visual embedding.

T_{cls}^{out}: *local* class-specific visual tokens, refined by *sample-specific visual context*.

Image-language context transfer



The benefit is two-fold:

- Transferring rich image-related language context from CLIP to the output text tokens
- Batch contrastive loss enhances the discriminative ability of the output text tokens across samples

$$\mathcal{L}_{contrast} = CrossEntropy(\mathbf{S}, \mathbf{I})$$

$\mathbf{S} \in \mathbb{R}^{B\times B}$ is the similarity matrix, $\mathbf{I} \in \mathbb{R}^{B\times B}$ is an identity matrix, B is the batch size.

Training objectives

Text-to-patch correlation maps:

$$\mathbf{C}_{t2p} = torch.matmul(\mathbf{T}_{txt}, \mathbf{T}_{pat})$$

$$\mathbf{y}_{t2p} = G(\mathbf{C}_{t2p})$$

$\mathbf{T}_{txt} \in \mathbb{R}^{C \times D}, \mathbf{T}_{pat} \in \mathbb{R}^{D \times HW}, \boldsymbol{C}_{t2p} \in \mathbb{R}^{C \times HW}$

Class-to-patch correlation maps:

$$\mathbf{C}_{c2p} = torch.matmul(\mathbf{T}_{cls}, \mathbf{T}_{pat})$$

$$\mathbf{y}_{c2p} = G(\mathbf{C}_{c2p})$$

$\mathbf{T}_{cls} \in \mathbb{R}^{C \times D}, \mathbf{T}_{pat} \in \mathbb{R}^{D \times HW}, \boldsymbol{C}_{c2p} \in \mathbb{R}^{C \times HW}$

Global Weighted Ranking Pooling:

For a class c,

$$G_c(\mathbf{C}) = \frac{1}{Z(d)} \sum_{j=1}^{HW} d^{j-1} \mathbf{C}^{r_j, c}$$

$$\mathbf{C}^{r_1,c} > \mathbf{C}^{r_2,c} > \cdots > \mathbf{C}^{r_{HW},c}$$

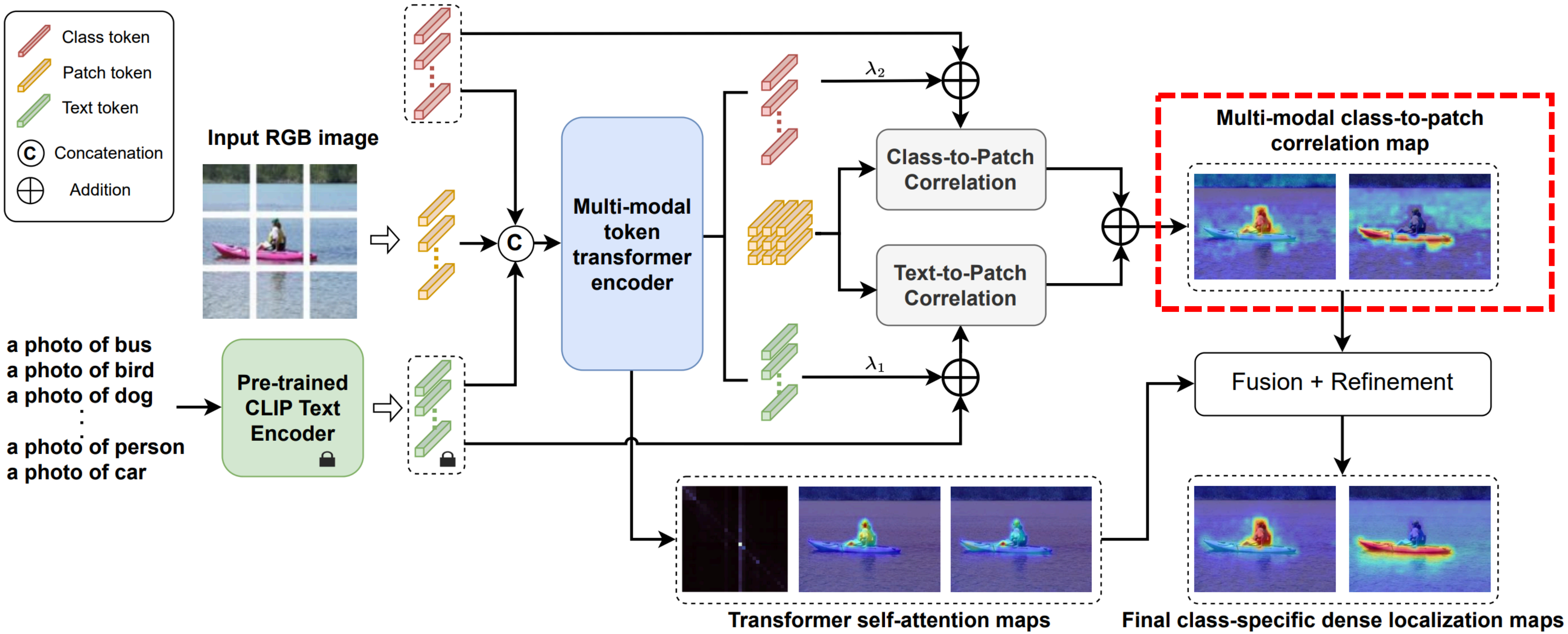
$$Z(d) = \sum_{j=1}^{HW} d^{j-1}$$

C is the correlation map; d is a decay parameter.

$$\mathcal{L}_{total} = \mathcal{L}_{MLSM}^{cls-token} + \mathcal{L}_{MLSM}^{t2p} + \mathcal{L}_{MLSM}^{c2p} + \mathcal{L}_{contrast}$$

MLSM: multi-label soft margin loss.

Class-specific dense localization inference



Global-local multi-modal class-specific tokens

Evaluation of the generated dense object localization on the train set of PASCAL VOC

Class representations	mIoU
Global class-specific visual tokens	62.7
Global multi-modal class-specific tokens	64.1
Local multi-modal class-specific tokens	63.3
Global-local multi-modal class-specific tokens	66.3

Image-language context transfer

Evaluation of the generated dense object localization on the train set of PASCAL VOC

Visual context	Image-language context: Prior knowledge	Image-language context: Regularization loss	mIoU
✗	-	-	64.1
✓	-	-	64.8
✓	CLIP caption embed.	L1	63.7
✓	CLIP caption embed.	Batch-contrast CE	65.1
✓	CLIP image embed.	Batch-contrast CE	**66.3**

Comparison with SoTA WSDOL methods

Multi-label dense localization

Method	Cls. Backbone	VOC	COCO
CAM (CVPR16) [48]	ResNet50	48.8	33.5†
SEAM (CVPR20) [34]	ResNet38	55.4	25.1‡
RIB (NeurIPS21) [18]	ResNet50	56.5	36.5
AdvCAM (CVPR21) [19]	ResNet38	55.6	37.2
CLIMS (CVPR22) [37]	ResNet50	56.6	-
SIPE (CVPR22) [4]	ResNet50	58.6	-
W-OoD (CVPR22) [21]	ResNet50	59.1	-
Du *et al.* (CVPR22) [8]	ResNet38	61.5	-
TS-CAM (ICCV21) [10]	ViT-small	41.3	-
MCTformer (CVPR22) [40]	ViT-small	61.7	-
MCTformer (CVPR22) [40]	ViT-base	62.3*	-
Ours	ViT-base	**66.3**	**40.9**

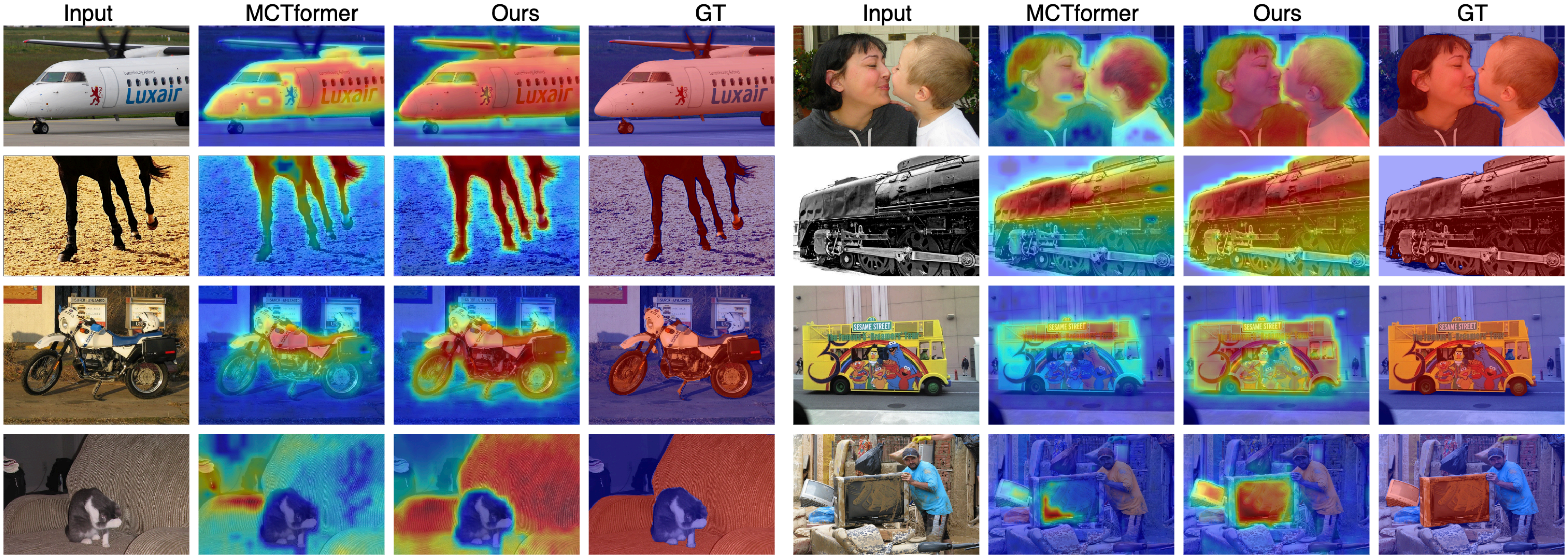
Single-label dense localization (OpenImages)

Method	Cls. backbone	pIoU	PxAP
CAM (CVPR16) [48]	ResNet50	43.0	58.2
HAS (ICCV17) [31]	ResNet50	41.9	55.1
ACoL (CVPR18) [47]	ResNet50	41.7	56.4
SPG (ECCV18) [46]	ResNet50	41.8	55.8
ADL (CVPR19) [6]	ResNet50	42.1	55.0
CutMix (ICCV19) [43]	ResNet50	42.7	57.6
PAS (ECCV20) [2]	ResNet50	-	60.9
IVR (ICCV21) [15]	ResNet50	-	58.9
Zhu *et al.* (CVPR22) [52]	ResNet50	49.7	65.4
CREAM (CVPR22) [38]	ResNet50	-	64.7
Zhu *et al.* (ECCV22) [51]	ResNet50	52.2	67.7
Ours	ViT-base	**57.6**	**73.3**

Comparison with SoTA WSSS methods

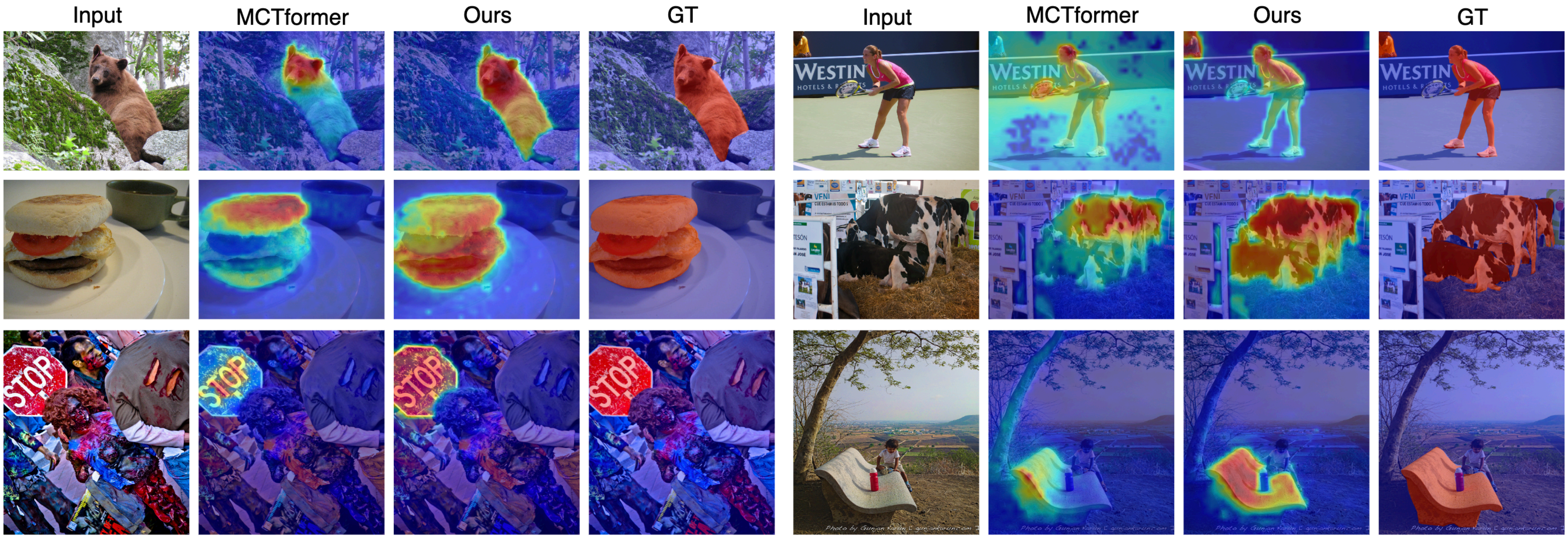
Method	Backbone	DeepLab version	Supervision	VOC		MS COCO
				Val	Test	Val
AuxSegNet (ICCV21)	ResNet38	V1	I+S	69.0	68.6	33.9
L2G (CVPR22)	ResNet38	V1	I+S	72.0	73.0	44.2
Kweon et al. (ICCV21)	ResNet38	V1	I	68.4	68.2	36.4
CDA (ICCV21)	ResNet38	V1	I	66.1	66.8	33.2
MCTformer (CVPR22)	ResNet38	V1	I	71.9	71.6	42.0
SIPE (CVPR 22)	ResNet38	V1	I	68.2	69.7	43.6
Yoon et al. (ECCV22)	ResNet38	V1	I	70.9	71.7	44.8
CLIMS (CVPR22)	ResNet101	V2	I+L	69.3	68.7	-
Ours	ResNet38	V1	I+L	72.2	72.2	45.9

Qualitative results on PASCAL VOC



MCTformer: Multi-class token transformer for weakly supervised semantic segmentation, CVPR 2022.

Qualitative results on MS COCO



MCTformer: Multi-class token transformer for weakly supervised semantic segmentation, CVPR 2022.

Qualitative results on OpenImages