OTSeg: Multi-prompt Sinkhorn Attention for Zero-Shot Semantic Segmentation

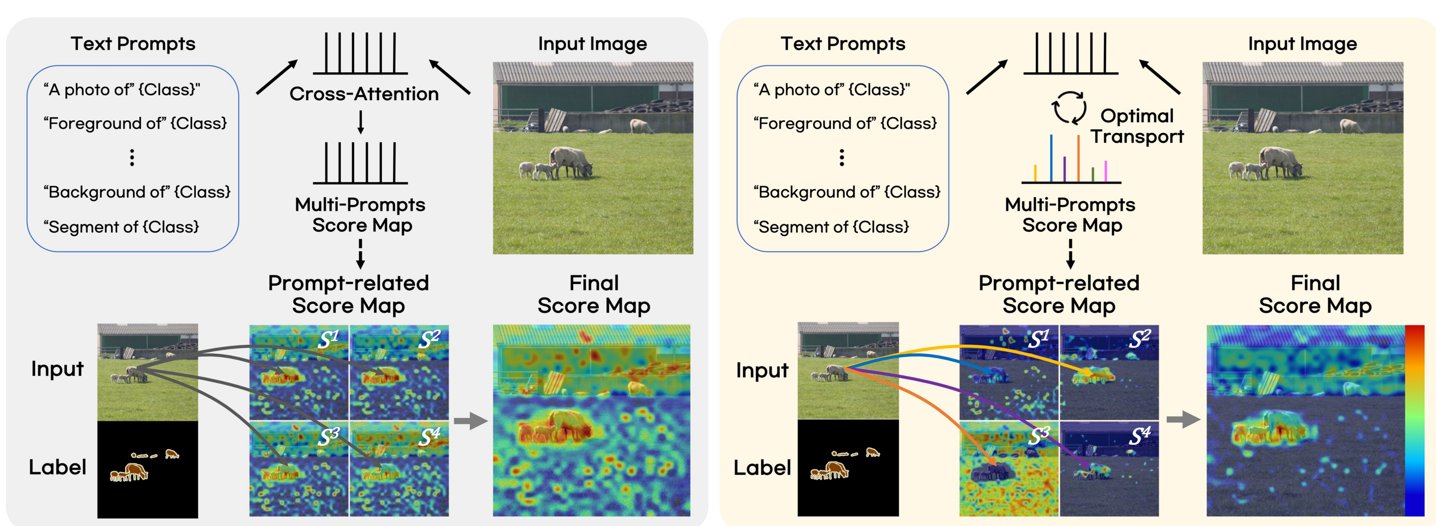
Kwanyoung Kim*[1], Yujin Oh*[2], and Jong Chul Ye[1]

[1]Korea Advanced Institute of Science and Technology (KAIST) [2]Massachusetts General Hospital (MGH) and Harvard Medical School

EUROPEAN CONFERENCE ON COMPUTER VISION

Introduction

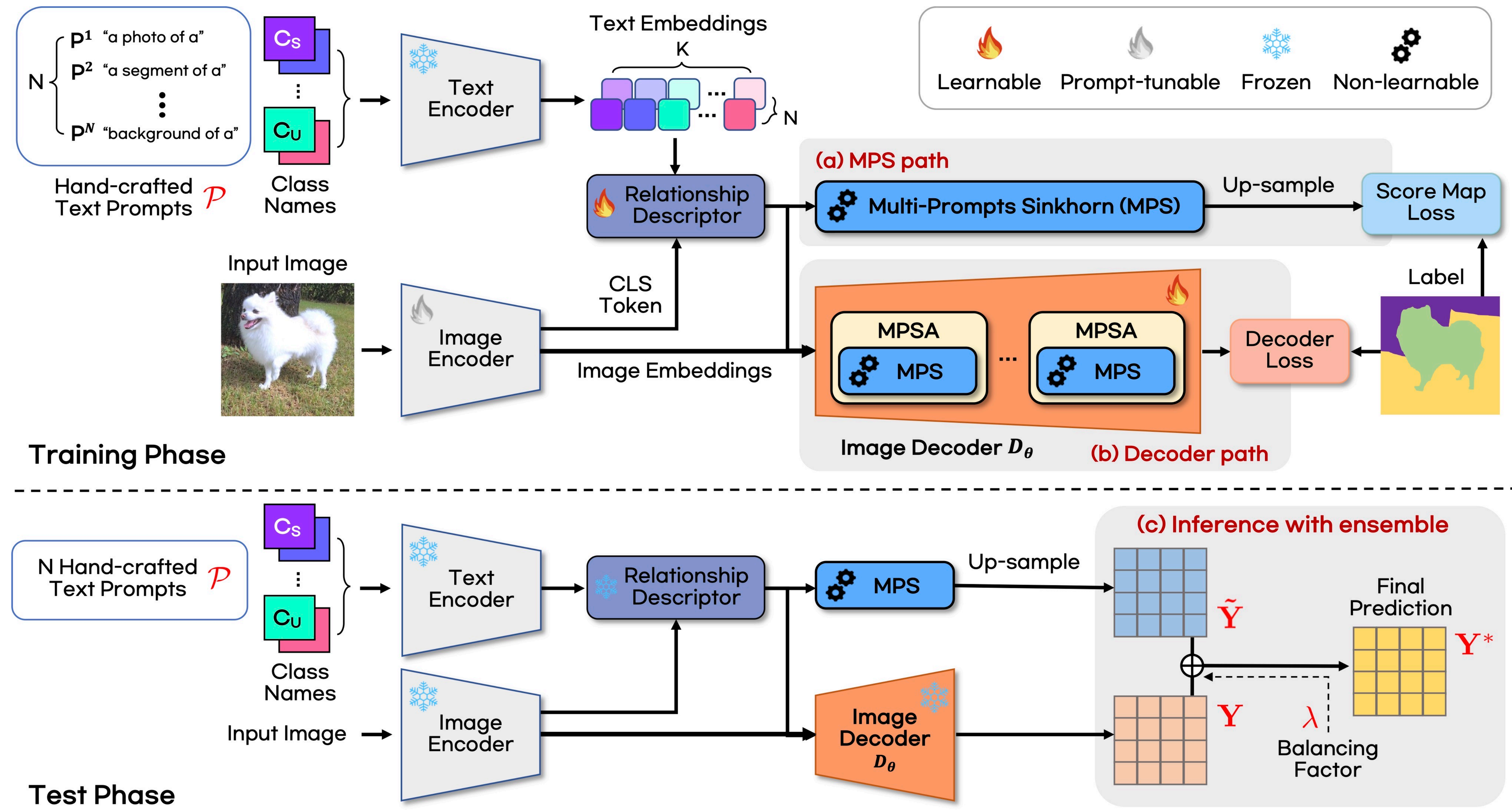
- Despite the success of CLIP transfer learning for zero-shot semantic segmentation (ZS3), challenges remain in closely aligning text embeddings with pixel embeddings.
- To address this issue, we propose OTSeg, a multimodal attention mechanism aimed at enhancing the effectiveness of matching multiple text prompts with corresponding pixel embeddings.
- Specifically, we propose Multi-Prompt Sinkhorn (MPS) based on the Optimal Transport, and along with its extension, Multi-Prompts Sinkhorn Attention (MPSA) which effectively replaces cross-attention mechanisms within Transformer.



(a) Default Cross-Attention (b) Multi-Prompts Sinkhorn Attention (MPSA)

Overall Pipeline of OTSeg+

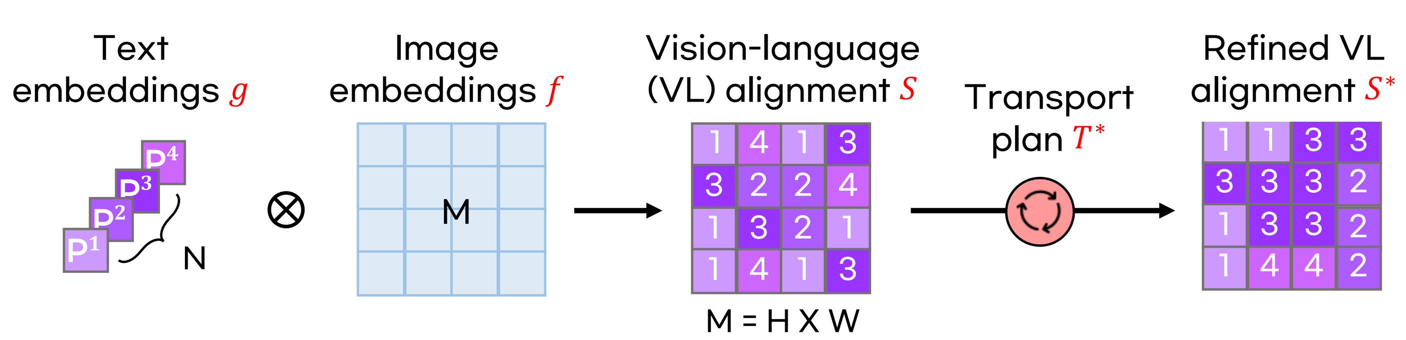
- OTSeg+ ensembles (a) MPS path output and (b) Decoder path output to yield final prediction.



Multi-Prompt Sinkhorn (MPS)

- Inspired by Sinkhorn algorithm, we propose MPS to allocate M-image pixel embeddings to N-text prompt embeddings.
- The discrete optimal transport plan $\boldsymbol{T}^*$ tries to maximize the Vision-language (VL) alignment $\boldsymbol{S}$ by minimizing the total cost $\boldsymbol{C}$.

$$\boldsymbol{S}^* = \mathrm{MPS}(\boldsymbol{S}) = \mathcal{M}(\boldsymbol{T}^* \odot \boldsymbol{S}),$$
$$\text{where } \boldsymbol{T}^* = \mathrm{Sinkhorn}\left(\frac{\boldsymbol{C}}{\epsilon}\right),\ \boldsymbol{C} := 1 - \boldsymbol{S}$$



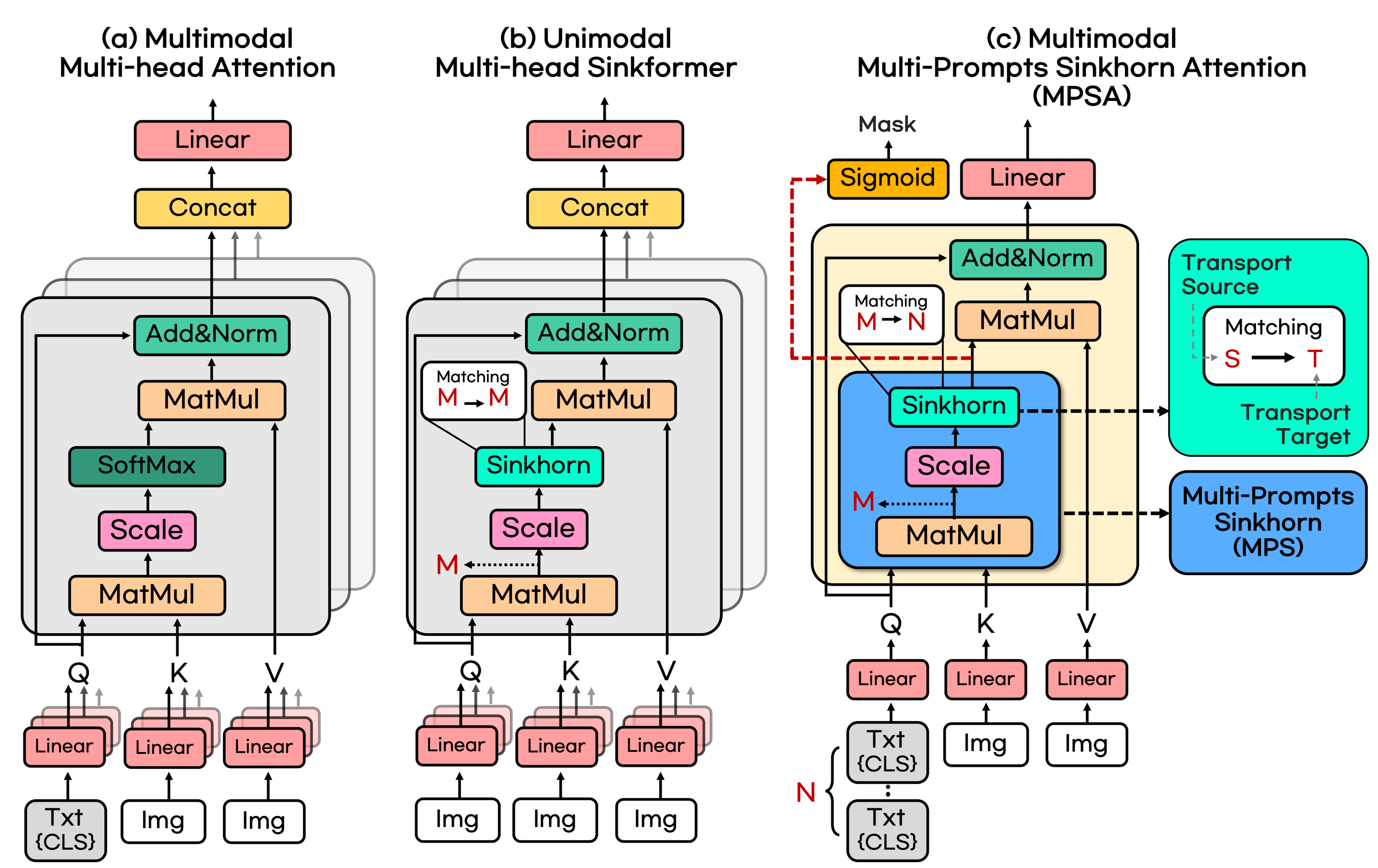
- The total cost $\boldsymbol{C}$ is the matrix multiplication of the text embeddings $\boldsymbol{g}$ and the image embeddings $\boldsymbol{f}$.
- Minimizing the total cost $\boldsymbol{C}$ corresponds to maximizing the VL alignment $\boldsymbol{S}$ to yield the refined VL alignment $\boldsymbol{S}^*$.

$$\boldsymbol{T}^* = \underset{\boldsymbol{T}\in\mathbb{R}^{M\times N}}{argmin} \sum_{i=1}^{M}\sum_{j=1}^{N} \boldsymbol{T}_{ij}\boldsymbol{C}_{ij} - \epsilon H(\boldsymbol{T}),$$
$$\text{where } \boldsymbol{C}_{ij} = 1 - \frac{f_i g_i^{\top}}{\|f_i\|_2\,\|g_i\|_2} = 1 - \boldsymbol{S}_{ij}$$

Multi-Prompt Sinkhorn Attention (MPSA)

- MPS can be extended into the cross-attention mechanism, referred as MPSA and integrated as a plugin module in each decoder layer.

$$\mathrm{Cross\text{-}Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{SoftMax}\left(\frac{\mathbf{QK}^{\top}}{\sqrt{d}}\right)\mathbf{V} \xrightarrow{\text{Replace}} \mathrm{MPSA}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{MPS}(\mathbf{QK}^{\top})\mathbf{V}$$

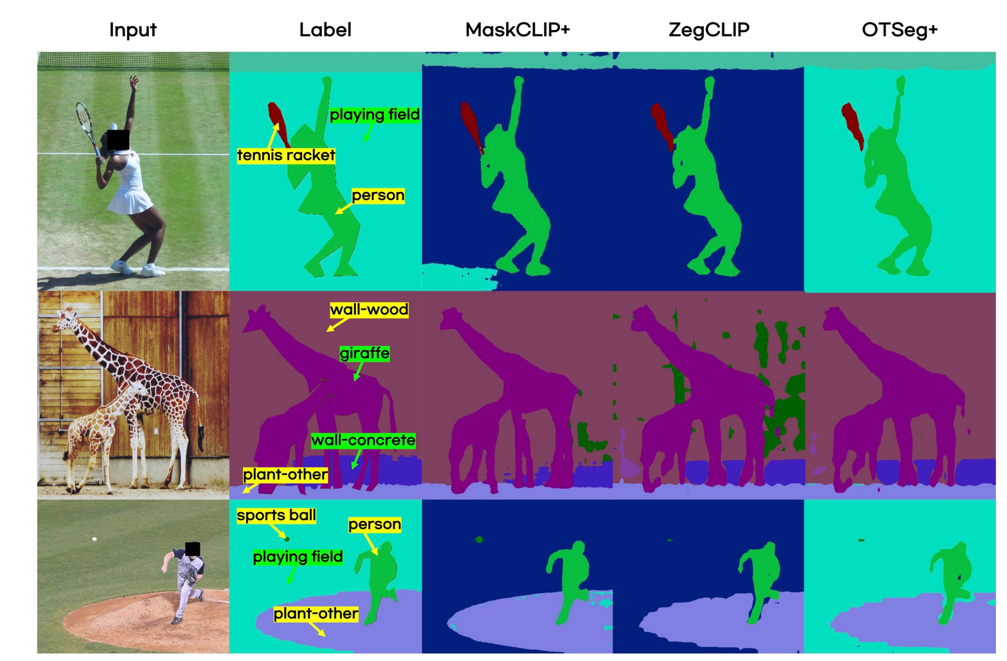


Comparison with Other Methods

- OTSeg+ achieves state-of-the-art performance across 3 benchmarks.

Methods	VOC 2012			PASCAL Context			COCO-Stuff164K		
	mIoU(U)	mIoU(S)	hIoU	mIoU(U)	mIoU(S)	hIoU	mIoU(U)	mIoU(S)	hIoU
Inductive setting									
ZegFormer [10]	63.6	86.4	73.3	-	-	-	33.2	36.6	34.8
Zsseg [35]	72.5	83.5	77.6	-	-	-	36.3	39.3	37.8
ZegCLIP [38]	77.8	91.9	84.3	54.6	46.0	49.9	41.4	40.2	40.8
OTSeg	78.1	92.1	84.5	56.7	53.0	54.8	41.4	**41.4**	41.4
OTSeg+	**81.6**	**93.3**	**87.1**	**60.4**	**55.2**	**57.7**	**41.8**	41.3	**41.5**
Transductive setting									
Zsseg [35]	78.1	79.2	79.3	-	-	-	43.6	39.6	41.5
MaskCLIP+ [37]	88.1	86.1	87.4	66.7	48.1	53.3	54.7	39.6	45.0
FreeSeg [25]	82.6	91.8	86.9	-	-	-	49.1	42.2	45.3
MVP-SEG+ [14]	87.4	89.0	88.0	67.5	48.7	54.0	55.8	39.9	45.5
ZegCLIP [38]	89.9	92.3	91.1	**68.5**	46.8	55.6	59.9	40.7	48.5
OTSeg	94.3	94.2	94.2	66.7	53.4	59.3	60.7	**41.8**	49.5
OTSeg+	**94.3**	**94.3**	**94.4**	67.0	**54.0**	**59.8**	**62.6**	41.4	**49.8**
Fully-supervised									
ZegCLIP [38]	90.9	92.4	91.6	**78.7**	46.5	56.9	63.2	40.7	49.6
OTSeg	94.4	94.0	94.2	78.1	55.2	64.7	**64.0**	**41.8**	**50.5**
OTSeg+	**95.0**	**94.1**	**94.6**	78.4	**54.5**	**65.5**	63.2	41.5	50.1

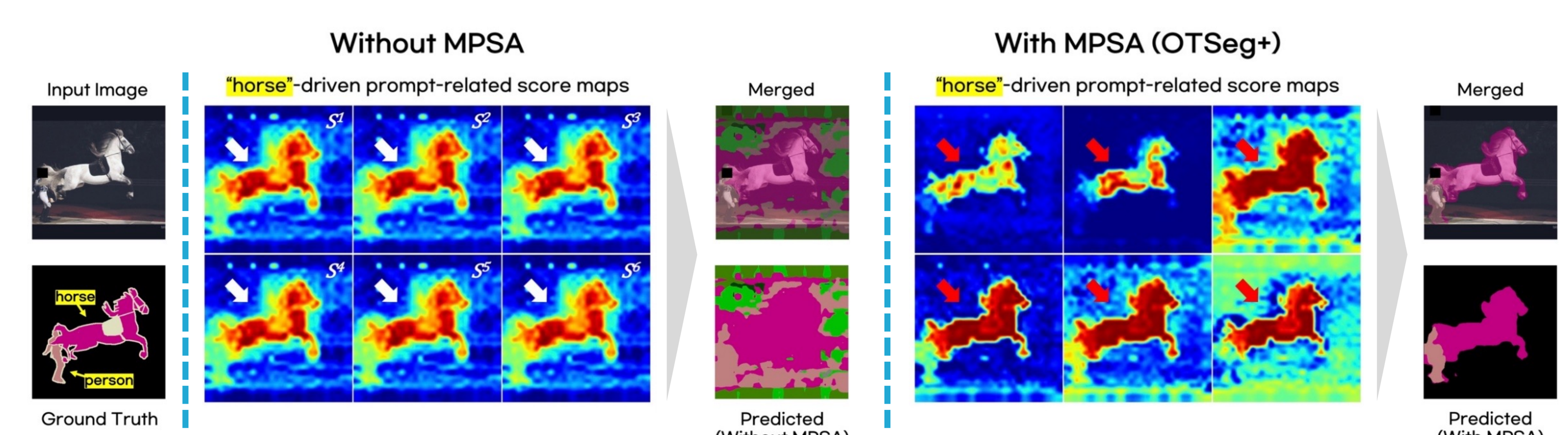
Source Code



Project

- OTSeg+ demonstrates efficiency in both computational cost and attention mechanisms by providing diversely dispersed score maps.

Method	# Parameter (M) ↓	GFLOPS ↓	FPS ↑
Zsseg	61.1	1916.7	4.2
ZegFormer	60.3	1829.3	6.8
ZegCLIP	**13.8**	**61.1**	**25.6**
OTSeg	**13.8**	$61.9_{-0.8}$	$23.6_{-2.0}$
OTSeg+	**13.8**	$61.9_{-0.8}$	$22.5_{-3.1}$



Conclusion

- We introduce OTSeg, a novel multimodal matching framework for ZS3.
- Our proposed component MPS and MPSA can propose the way for new directions for future researches in multimodal alignment and zero-shot learning requiring multi-conceptual semantic understandings of vision.