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

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

[1] Graduate School of AI, Korea Advanced Institute of Science and Technology (KAIST), South Korea
[2] Department of Radiology, Massachusetts General Hospital (MGH) and Harvard Medical School, Boston, MA, USA
*Equal contribution.

Abstract. The recent success of CLIP has demonstrated promising results in zero-shot semantic segmentation by transferring muiltimodal knowledge to pixel-level classification. However, leveraging pre-trained CLIP knowledge to closely align text embeddings with pixel embeddings still has limitations in existing approaches. To address this issue, we propose OTSeg, a novel multimodal attention mechanism aimed at enhancing the potential of multiple text prompts for matching associated pixel embeddings. We first propose Multi-Prompts Sinkhorn (MPS) based on the Optimal Transport (OT) algorithm, which leads multiple text prompts to selectively focus on various semantic features within image pixels. Moreover, inspired by the success of Sinkformers in unimodal settings, we introduce the extension of MPS, called Multi-Prompts Sinkhorn Attention (MPSA), which effectively replaces cross-attention mechanisms within Transformer framework in multimodal settings. Through extensive experiments, we demonstrate that OTSeg achieves state-of-the-art (SOTA) performance with significant gains on Zero-Shot Semantic Segmentation (ZS3) tasks across three benchmark datasets. We release our source code at https://github.com/cubeyoung/OTSeg.

Keywords: Multimodal · Sinkhorn · Cross-Attention · Segmentation

1 Introduction

Transformer's attention mechanism has demonstrated its remarkable performance across various tasks, establishing itself as a universal backbone structure in foundational models [4, 9, 26, 30]. However, recent advancements cast doubt on the optimality of traditional transformer structures that rely solely on consecutive self-attention layers. Among widespread efforts to enhance transformer performance [16, 21, 22], one notable endeavor is Sinkformer [28], where the SoftMax normalization of self-attention mechanism within transformer is simply replaced by the optimal transport (OT) algorithm, yielding a doubly stochastic attention that improves model performance in each vision and linguistic task.

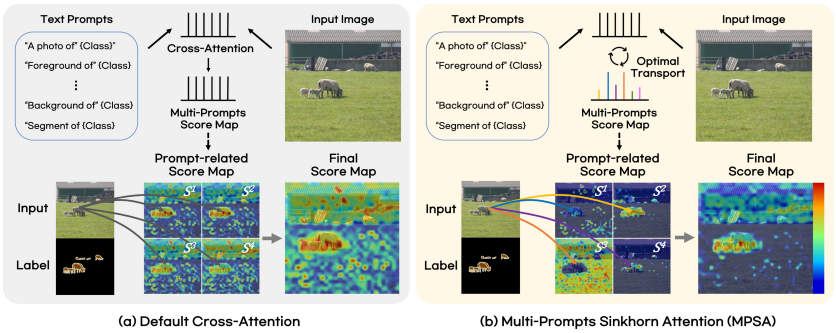

Fig. 1: Visualization of proposed Multi-Prompts Sinkhorn Atttention (MPSA) for text-driven semantic segmentation. (a) Without MPSA, all the text prompt-related score maps S^i are cohered. (b) With MPSA, each S^i selectively focuses on different semantic attributes, resulting the final score map effectively attends to the target object.

While Sinkformer is exclusively designed for unimodal settings, we are interested in its more appropriate applicability for multimodal alignment, particularly in the context of text-driven semantic segmentation tasks. In text-driven semantic segmentation, zero-shot semantic segmentation (ZS3) [2] represents a label-efficient approach. A key factor contributing to the recent advancements in ZS3 solutions [10, 27, 34, 36, 37] is application of pre-trained Vision-Language Model (VLM), such as CLIP [26]. However, care should be taken when transferring VLM, as the pre-trained knowledge is not optimized for pixel-wise dense alignment, since VLM has been trained with a variety of image-text pairs through contrastive learning. One naive solution can be fine-tuning the VLM tailored for desired pixel-level prediction by leveraging an ensemble technique driven by multiple text prompts [1]. However, as shown in Fig. 1(a), this naive approach is still problematic because the introduced multiple text prompts and image pixels are passively aligned, which leads each text prompt-driven pixel-level prediction to be cohered to each other.

To address the limited potential of multiple text prompts, we introduce a pixel-text alignment operator based on the OT algorithm, namely, Multi-Prompts Sinkhorn (MPS). Furthermore, as like the aforementioned Sinkformer improved the traditional transformer by introducing doubly stochastic self-attention, we have discovered that the MPS algorithm can serve as an optimal replacement for the SoftMax normalization of cross-attention mechanism within multimodal transformer. This leads to the development of a novel Multi-Prompts Sinkhorn Attention (MPSA) module, which composes our proposed OTSeg. As shown in Fig. 1(b), we empirically find that our OTSeg enhances the diversity of text prompt-driven pixel-level predictions by selectively focusing on various semantic features. Consequently, the optimally ensembled final prediction effectively focuses on the target object, leading to the achievement of the state-of-the-art (SOTA) performance in ZS3 tasks. Our contributions can be summarized as:

– We introduce a novel OTSeg that yields diverse pixel-level predictions driven by multiple text prompts, resulting in improved multimodal alignment.

– Through extensive experiments on three benchmark datasets, OTSeg demonstrates its superiority on ZS3 tasks by achieving SOTA performance.

2 Related Work

2.1 Zero-shot & Open-vocabulary Semantic Segmentation

Semantic segmentation is a core computer vision task to densely analyze visual context. Recent success of CLIP [26] accelerates the advancement of language-driven semantic segmentation by utilizing pre-trained knowledge of VLM [17, 18, 33]. However, since this dense prediction task requires a labor-intensive pixel-level annotation, there arise a label-imbalance issue, *i.e.,* not all the categories are annotated in the training dataset. Zero-shot semantic segmentation (ZS3) solves this label-imbalance problem by generalizing labeled (seen) class knowledge to predict new (unseen) class information [2]. MaskCLIP+ [37] introduces a ZS3 method by simply extracting the text-driven visual features from the CLIP image encoder. ZegCLIP [37] successfully bridges the performance gap between the seen and unseen classes by adapting a visual prompt tuning technique instead of fine-tuning the frozen CLIP image encoder. Recently, FreeSeg [25] and MVP-SEG+ [14] introduce text prompt-driven method for realizing open-vocabulary segmentation. In specific, MVP-SEG+ employs orthogonal constraint loss (OCL) to each prompt to exploit CLIP features on different object parts.

ZS3 can be performed by either inductive or transductive settings. Compared to inductive ZS3 where class names and pixel-level annotations of unseen classes are both unavailable during training [10], a newly introduced transductive setting boosts the ZS3 performance by utilizing unseen class names and self-generated pseudo labels guided by the model itself during training [13, 23, 34, 36, 37]. Open-vocabulary settings simply extend the concept of inductive ZS3 settings applied for cross-domain datasets. In this study, we apply our proposed OTSeg for both the inductive and transductive settings, and we further demonstrate our OTSeg performance on various cross-dataset settings.

2.2 Optimal Transport

Optimal transport (OT) is a general mathematical framework to evaluate correspondences between two distributions. Thanks to the luminous property of distribution matching, the optimal transport has received great attention and proven its generalization capability in various computer vision tasks, such as domain adaptation [12], semantic correspondence problem [19], graph matching [31, 32], and cross-domain alignment [6], etc. Among various methods, Sinkhorn algorithm can efficiently solve the OT problem through entropy-regularization [8], and it can be directly applied to deep learning frameworks thanks to the extension of Envelop Theorem [24]. In the context of computer vision tasks, prompt learning with optimal transport (PLOT) [5] optimizes the OT distance to align visual and text features by the Sinkhorn given trainable multiple text prompts for few-shot

image-level prediction tasks. The most related work to ours is Sinkformer [28], where SoftMax layers within self-attention transformer is replaced by Sinkhorn algorithm, resulting in enhanced accuracy in each vision and natural language processing task. The Sinkformer has inspired us to propose the Sinkhorn algorithm as an ideal fit for multimodal alignment, leading us to apply OT to further boost the performance of the cross-attention mechanism for ZS3 tasks.

3 Preliminary

3.1 Optimal Transport Problem and Sinkhorn

Optimal transport aims to minimize the transport distance between two probability distributions. In this paper, we only consider discrete distribution which is closely related to our framework. We assume the feature vector F, G are defined as $F = \{\boldsymbol{f}_i\}_{i=1}^{M}$ and $G = \{\boldsymbol{g}_j\}_{j=1}^{N}$ and discrete empirical distributions $\mathbf{u}$ and $\mathbf{v}$ that are defined on probability space $\mathcal{F}, \mathcal{G} \in \Omega$, respectively, as follows:

$$\mathbf{u} = \sum_{i=1}^{M} \mu_i \delta_{\boldsymbol{f}_i}, \quad \mathbf{v} = \sum_{j=1}^{N} \nu_j \delta_{\boldsymbol{g}_j}, \tag{1}$$

where $\delta_{\boldsymbol{f}}$ and $\delta_{\boldsymbol{g}}$ denote Dirac functions centered on $\boldsymbol{f}$ and $\boldsymbol{g}$, respectively, M and N denote the dimension of the empirical distribution. The weight vectors $\boldsymbol{\mu} = \{\mu_i\}_{i=1}^{M}$ and $\boldsymbol{\nu} = \{\nu_j\}_{j=1}^{N}$ belong to the M and N-dimensional simplex, respectively, *i.e.*, $\sum_{i=1}^{M} \mu_i = 1$, $\sum_{j=1}^{N} \nu_j = 1$. The discrete optimal transport problem can be then formulated as:

$$\begin{aligned} \mathbf{T}^* = & \underset{\mathbf{T} \in \mathbb{R}^{MXN}}{\arg\min} \sum_{i=1}^{M} \sum_{j=1}^{N} \mathbf{T}_{ij} \mathbf{C}_{ij} \\ \text{s.t.} \quad & \mathbf{T}\mathbf{1}^N = \boldsymbol{\mu}, \quad \mathbf{T}^\top \mathbf{1}^M = \boldsymbol{\nu}. \end{aligned} \tag{2}$$

Here, $\mathbf{T}^*$ is called the optimal transport plan, which is learned to minimize the total distance between the two probability vectors. $\mathbf{C}$ is the cost matrix which represents the distance between $\boldsymbol{f}_i$ and $\boldsymbol{g}_j$, *e.g.*, the cosine distance $\mathbf{C}_{ij} = 1 - \frac{\mathbf{f}_i \mathbf{g}_j^\top}{\|\mathbf{f}_i\|_2 \|\mathbf{g}_j\|_2}$, and $\mathbf{1}^M$ refers to the M-dimensional vector with ones.

However, solving the problem Eq. (2) costs $O(n^3 \log n)$-complexity (n proportional to M and N), which is time-consuming. This issue can be efficiently solved by the entropy-regularization of the objective through the Sinkhorn-Knopp (or simply Sinkhorn) algorithm [8]. In Sinkhorn algorithm, the optimization problem is reformulated as:

$$\begin{aligned} \mathbf{T}^* = & \underset{\mathbf{T} \in \mathbb{R}^{MXN}}{\arg\min} \sum_{i=1}^{M} \sum_{j=1}^{N} \mathbf{T}_{ij} \mathbf{C}_{ij} - \epsilon H(\mathbf{T}) \\ & \text{s.t.} \quad \mathbf{T}\mathbf{1}^N = \boldsymbol{\mu}, \quad \mathbf{T}^\top \mathbf{1}^M = \boldsymbol{\nu}. \end{aligned} \tag{3}$$

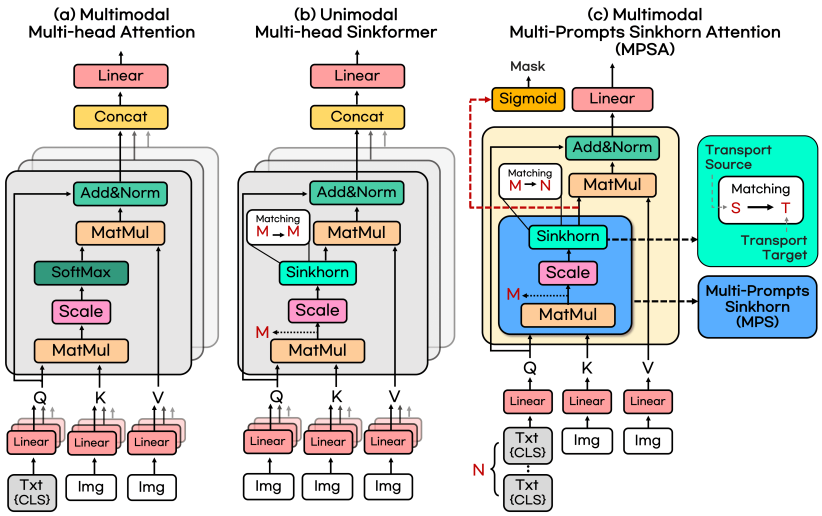

Fig. 2: Comparison of attention mechanism variants. (a) Cross-attention mechanism for multimodal settings. (b) Sinkformer self-attention mechanism for unimodal settings. (c) Our proposed Muti-Prompt Sinkhorn Attention (MPSA) for multimodal settings, which aims to optimally transport image pixel (M) to multiple text prompts (N).

where $H(\mathbf{T}) = \sum_{ij} \mathbf{T}_{ij} \log \mathbf{T}_{ij}$ and $\epsilon > 0$ is the regularization parameter. For the problem Eq. (3), we have an optimization solution when $t \to \infty$ as follow:

$$\mathbf{T}^* = \text{diag}(\mathbf{a}^t) \exp(-\mathbf{C}/\epsilon) \text{diag}(\mathbf{b}^t) \tag{4}$$

where t is the iteration and $\mathbf{a}^t = \boldsymbol{\mu} / \exp(-\mathbf{C}/\epsilon)\mathbf{b}^{t-1}$ and $\mathbf{b}^t = \boldsymbol{\nu} / \exp(-\mathbf{C}/\epsilon)\mathbf{a}^t$, with the initialization on $\mathbf{b}^0 = \mathbf{1}$. To stabilize the iterative computations, we adopt the log scaling version of Sinkhorn optimization [29].

3.2 Self-Attention and Sinkformer

The transformer model employs a self-attention mechanism when given a sequence of length $\mathbf{X} = [x_1, x_2, \cdots, x_n]$ embedded in d dimensions:

$$\text{Self-Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \texttt{SoftMax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V} \tag{5}$$

$$\text{where} \quad \mathbf{Q} = \phi_q(\mathbf{X}),\ \mathbf{K} = \phi_k(\mathbf{X}),\ \mathbf{V} = \phi_v(\mathbf{X}), \tag{6}$$

where $\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{M \times d}$, and $\mathbf{V} \in \mathbb{R}^{d \times d}$ denotes query, key and value matrices, respectively. ϕ_q, ϕ_k and ϕ_v represent the linear layer of each matrix. In Eq.(6), the attention matrix is normalized using the `SoftMax` operator. In Sinkformer, it is demonstrated that the initial iteration in the Sinkhorn algorithm is identical to the softmax operation. This observation leds to the replacement of `SoftMax` with the Sinkhorn algorithm in self-attention to ensure double stochasticity. For

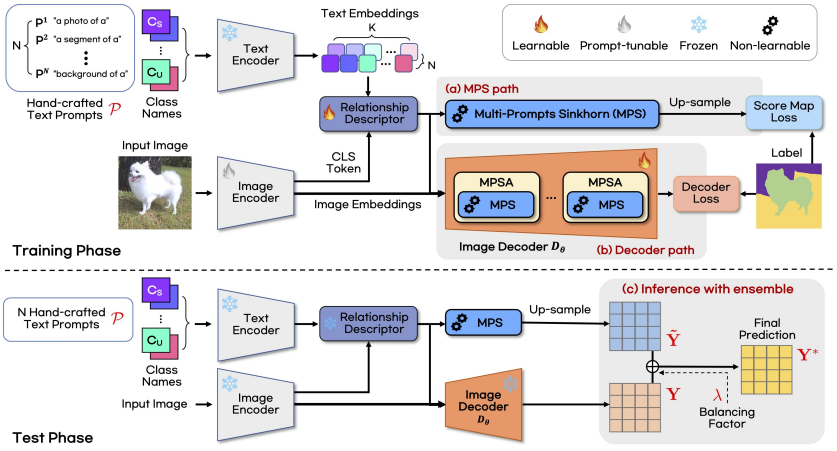

Fig. 3: Overview of OTSeg for zero-shot semantic segmentation. (a) MPS path refines the score map using the `MPS` algorithm. (b) Decoder path involves the decoder output, which integrates the Multi-Prompts Sinkhorn Attention (MPSA) predictions. (c) During inference, OTSeg ensembles predictions from both paths with a balancing factor λ.

simplicity, Sinkhorn's algorithm is seamlessly integrated into the self-attention modules as follow:

$$\text{Sinkformer Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \texttt{Sinkhorn}\left(\frac{\mathbf{C}}{\sqrt{d}}\right)\mathbf{V},\ \mathbf{C} = \mathbf{1} - \mathbf{Q}\mathbf{K}^\top \tag{7}$$

where $\texttt{Sinkhorn}(\cdot)$ is the Sinkhorn operator, which reduces the cost $\mathbf{C}$ by utilizing Eq. (4). The scale factor ϵ is replaced by $\sqrt{d}$. Despite the extension, Sinkformer is still limited in unimodal settings such as the self-attention within an image modality (M pixels→M pixels), as illustrated in Fig. 2(b).

4 Methods

In this section, we present a method for performing zero-shot segmentation tasks using our proposed OTSeg framework for multimodal setting (M pixels→N text prompts), as illustrated in Fig. 2(c). To clarify, we define several notations within our OTSeg framework: a pair of frozen CLIP text encoder and tunable image encoder, the relationship descriptor, multiple hand-crafted text prompts $\mathcal{P}$, and trainable decoder D_θ. Furthermore, we propose three fundamental components of OTSeg: (a) Multi-Prompts Sinkhron (MPS), (b) Multi-Prompts Sinkhorn Attention (MPSA) and (c) Inference with ensemble.x In Sec. 4.1, we introduce the concept of multiple prompts-guided text embeddings and explain how they are processed. In Sec. 4.2, we provide a detailed account of our OTSeg, including aforementioned three key components. Lastly, in Sec. 4.3, we describe the training procedure for our method and introduce the associated loss functions.

4.1 Multiple Prompts-guided Text Embeddings

To effectively transfer CLIP's pre-trained knowledge, we adopt the frozen text encoder with multiple hand-crafted text prompts and a tunable image encoder, as shown in Fig. 3. The multiple text prompts are created as a set of N text prompts denoted as $\mathcal{P} = \{\mathbf{P}^i|_{i=1}^N\}$. Each text prompt $\mathbf{P}^i$ can be defined as $\mathbf{P}^i = [P_1^i, P_2^i, \cdots, P_l^i]$, where l represents the length of the context tokens. These text prompts are then consistently added in front of K tokenized class names, forming a set denoted as $\mathcal{T} = \{\{\mathcal{P}, \mathbf{c}^k\}|_{k=1}^K\}$. Note that the same text prompts $\mathcal{P}$ are shared across all class names. Here, $\{\mathbf{c}^k|_{k=1}^K\}$ represents the word embeddings of each class name, drawn from a larger set $\mathcal{C}$. Then, the set $\mathcal{T}$ is inputted to the frozen CLIP text encoder, resulting in the text embeddings $\mathbf{h}_{\text{txt}} \in \mathbb{R}^{KN \times D}$, where D represents the embedding dimension. We utilize the relationship descriptor following an approach introduced in [37] to yield the refined text embedding $\tilde{\mathbf{h}}_{\text{txt}} \in \mathbb{R}^{KN \times D}$. The detail of relationship descriptor are deferred in Appendix **??**.

4.2 Optimal Transformer for Zero-Shot Segmentation

(a) Multi-Prompts Sinkhorn (MPS) Now, when we input an image through the tunable CLIP image encoder, the model yields pixel embedding $\mathbf{h}_{\text{img}} \in \mathbb{R}^{M \times D}$, where $M = H \times W$ corresponds to the product of the height H and width W. Given the text embeddings and pixel embeddings, the text-pixel aligned score map can be formulated as follow:

$$\mathbf{S} = \tilde{\mathbf{h}}_{\text{txt}} \mathbf{h}_{\text{img}}{}^{\top}, \tag{8}$$

where the superscript $^{\top}$ refers to the transpose operation, both $\tilde{\mathbf{h}}_{\text{txt}}$ and $\mathbf{h}_{\text{img}}$ are $\mathcal{L}_2$ normalized along the embedding dimension, and the score map $\mathbf{S} \in \mathbb{R}^{HW \times KN}$ undergoes further refinement.

To transport the distribution of the multiple text prompts to pixel distribution, we first define the total cost matrix $\mathbf{C}$ in Eq. (3) using the text-pixel aligned score map $\mathbf{S}$ in Eq. (8). Specifically, we set $\mathbf{C} := \mathbf{1} - \mathbf{S}$, where $\mathbf{C} \in \mathbb{R}^{HW \times KN}$ denotes the cost matrix. Given the cost matrix $\mathbf{C}$, the goal of `MPS` is to obtain the corresponding optimal transport plan $\mathbf{T}^*$ as given in Eq. (4), which aims to allocate each of the M image pixels to the N text-prompts, thereby allowing multiple text prompts to be associated with each pixel. Thus, $\mathbf{T}^*$ serves as a mapping matrix that maximizes the cosine similarity between multimodal embeddings, as outlined as **??** in **??**. Therefore, we formulate the refined score map through `MPS` algorithm:

$$\mathbf{S}^* = \texttt{MPS}(\mathbf{S}) = \mathcal{M}(\mathbf{T}^* \odot \mathbf{S}) \tag{9}$$

$$\text{where } \mathbf{T}^* = \texttt{Sinkhorn}\left(\frac{\mathbf{C}}{\epsilon}\right), \ \mathbf{C} := \mathbf{1} - \mathbf{S} \tag{10}$$

where $\mathcal{M}\colon \mathbb{R}^{HW \times KN} \to \mathbb{R}^{HW \times K}$ is the operation which first reshapes $\mathbb{R}^{HW \times KN} \to \mathbb{R}^{HW \times K \times N}$ and performs summation of all the score maps along the N dimension,

ϵ denotes the scaling hyper-parameter. The refined score map $\mathbf{S}^* \in \mathbb{R}^{HW \times K}$ by adapting the transport plan $\mathbf{T}^*$ can be served as a stand-alone logit for segmentation mask as outlined in Sec. 4.2 (c). Note that the transport plan $\mathbf{T}^*$ in Eq. (4) only contains matrix multiplication and exponential operation, thus `MPS` algorithm is fully differentiable and the gradients can be back-propagated throughout the entire neural network.

(b) Muti-Prompts Sinkhorn Attention (MPSA) In the multimodal settings for text-driven semantic segmentation tasks, we can formulate cross-attention between pixel embeddings and classnames-driven text embeddings with multiple prompts:

$$\text{Cross-Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \texttt{SoftMax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V} \tag{11}$$

$$\text{where} \mathbf{Q} = \phi_q(\tilde{\mathbf{h}}_{\text{txt}}),\ \mathbf{K} = \phi_k(\mathbf{h}_{\text{img}}), \mathbf{V} = \phi_v(\mathbf{h}_{\text{img}}), \tag{12}$$

where $\mathbf{Q} \in \mathbb{R}^{KN \times D}$, and $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{M \times D}$ denotes query, key and value matrices, respectively, and ϕ denotes linear projection for each query, key, value. Instead of applying the `MPS` algorithm solely on the score map, we empirically find that `MPS` can be extended into the cross-attention mechanism and seamlessly integrated as a plugin module in each decoder layer. Similar to the cross-attention mechanism Eq. (11), we define our proposed Multi-Prompts Sinkhorn Attention as follows:

$$\text{Multi-Prompts Sinkhorn Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \texttt{MPS}\left(\mathbf{Q}\mathbf{K}^\top\right)\mathbf{V}. \tag{13}$$

In our proposed MPSA, instead of conventional multiple head dimensions, we have N multiple text-prompts dimensions, as shown in Fig. 2(c). In our decoder, comprising three layers of cross-attention transformer as illustrated in Fig. 3, each multi-head attention module is replaced by our proposed Multi-Prompts Sinkhorn Attention (MPSA) module. In the decoder, a semantic mask is calculated by taking the intermediate product of MPSA, denoted by the `MPS` operation, followed by the `Sigmoid` function:

$$\texttt{Mask} = \texttt{Sigmoid}\left(\texttt{MPS}(\mathbf{Q}\mathbf{K}^\top)\right). \tag{14}$$

where `Mask` denotes the semantic mask in the final decoder layer as shown in Fig. 2(c), while $\mathbf{Q}$ and $\mathbf{K}$ refer to the query and key matrices from the preceding layer, respectively.

Then, we can obtain the final decoder output by applying the up-sampling operator as follow:

$$\mathbf{Y} = \mathcal{U}(\texttt{Mask}),\ \in \mathbb{R}^{H_I W_I \times K} \tag{15}$$

where $\mathbf{Y} \in \mathbb{R}^{H_I W_I \times K}$ is the final output of decoder which is integrated in our MPSA module, $\mathcal{U} : \mathbb{R}^{HW \times K} \rightarrow \mathbb{R}^{H_I W_I \times K}$ is the up-sampling operator ($HW < H_I W_I$), where H_I and W_I are height and width of the input image,

respectively. When the prediction is calculated from Eq. (15) using decoder, we refer it as OTSeg.

(c) Inference with ensemble Rather than solely relying on the prediction in Eq. (15), we further utilize the refined score map $\mathbf{S}^*$ in Eq. (9) to boost the segmentation performance. For this purpose, $\mathbf{S}^*$ is up-sampled to match the original image size as follows:

$$\tilde{\mathbf{Y}} = \mathcal{U}(\mathbf{S}^*). \tag{16}$$

where $\tilde{\mathbf{Y}} \in \mathbb{R}^{H_I W_I \times K}$ is the prediction of the refined score map. In order to synergistically exploit the collective knowledge derived from the learnable decoder in $\mathbf{Y}$ and information encapsulated in $\tilde{\mathbf{Y}}$, we formulate the final segmentation output $\mathbf{Y}^*$ as follows:

$$\mathbf{Y}^* = \lambda \cdot \mathbf{Y} + (1 - \lambda) \cdot \tilde{\mathbf{Y}} \tag{17}$$

where balance factor $\lambda \in [0, 1]$ denotes the hyper-parameter for controlling balance between $\mathbf{Y}$ and $\tilde{\mathbf{Y}}$, and set to 0.5 through component analysis in Appendix. We refer to this ensembled approach as OTSeg+.

4.3 Loss Function

In this work, we combine two different losses following previous methods as follows:

$$\mathcal{L}_{\text{seg}} = \lambda_{\text{fc}}\mathcal{L}_{\text{fc}} + \lambda_{\text{dc}}\mathcal{L}_{\text{dc}}, \;\; \mathcal{L}_{\text{tot}}(\Theta, \theta) = \mathcal{L}_{\text{seg}}(\mathbf{Y}, \mathbf{Y}^{\text{gt}}; \Theta, \theta) + \mathcal{L}_{\text{seg}}(\tilde{\mathbf{Y}}, \mathbf{Y}^{\text{gt}}; \Theta) \tag{18}$$

where $\Theta = [E_{\text{img}}, \mathcal{R}_\psi]$ contains tunable image encoder E_{img} and the linear layer $\mathcal{R}_\psi$, $\mathcal{L}_{\text{seg}}$ denotes the segmentation loss combining different two losses, $\mathcal{L}_{\text{fc}}$ and $\mathcal{L}_{\text{dc}}$ are the focal loss, and the dice loss, with λ_{fc}, and λ_{dc} as corresponding hyper-parameters, respectively, and $\mathbf{Y}^{\text{gt}} \in \mathbb{R}^{H_I W_I \times K}$ is the ground-truth label. The details of the loss function are described in **??**.

5 Experiments

5.1 Dataset and Evaluation Metric

Dataset A primary goal of ZS3 is to segment objects belong to both seen classes $\mathcal{C}_S$ and unseen classes $\mathcal{C}_U$, *i.e.*, $\mathcal{C} = \mathcal{C}_S \cup \mathcal{C}_U$, where $\mathcal{C}_S \cap \mathcal{C}_U = \emptyset$. For fair comparison with previous methods [2,10,34,36,37], we follow the identical protocol of dividing $\mathcal{C}_S$ and $\mathcal{C}_U$ for each dataset. To evaluate the effectiveness of our OTSeg, we carry out extensive experiments on three challenging datasets: VOC 2012 [11], PASCAL Context [20], and COCO-Stuff164K [3]. The dataset details are described in **??**.

Evaluation Metric We measure the mean of class-wise intersection over union (mIoU) on both seen and unseen classes, indicated as mIoU(S) and mIoU(U), respectively. We adopt the harmonic mean IoU (hIoU) of seen and unseen classes as a primary metric. More details are deferred to **??**.

Table 1: Quantitative comparison of zero-shot semantic segmentation performance with baseline methods. The **bold indicates the best performance.**

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 [34]	72.5	83.5	77.6	-	-	-	36.3	39.3	37.8
ZegCLIP [37]	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 [34]	78.1	79.2	79.3	-	-	-	43.6	39.6	41.5
MaskCLIP+ [36]	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 [37]	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 [37]	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

5.2 Implementation Details

We implemented the proposed method using the open-source toolbox MMSegmentation [7]. The algorithm was executed on up to 8 NVIDIA A100 GPUs with a batch size of 16. We utilized the pre-trained CLIP ViT-B/16 model[3] as the backbone VLM for all experiments. We fine-tuned the CLIP image encoder module by employing visual prompt tuning (VPT) approaches [15], while keeping the CLIP text encoder module frozen. The number of multiple text prompts was set to $N = 6$ for VOC 2012 and $N = 8$ for PASCAL Context and COCO-Stuff164K datasets. For the image decoder, we adopted a lightweight transformer consisting of three layers, with the original multi-head attention replaced by our MPSA. The optimizer was set to AdamW with a specific training schedule for each dataset.

5.3 Comparison in Zero-Shot Settings

Quantitative zero-shot segmentation results are presented in Tab. 1. We find that our proposed OTSeg outperforms the previous SOTA in both inductive and transductive settings. It implies that our proposed MPSA module decoder effectively enhances segmentation performance. Furthermore, we observe that our OTSeg+ achieves the best performance on all datasets and demonstrates the effectiveness of our ensemble strategy. Fig. 4 shows the qualitative zero-shot segmentation performance of our OTSeg+ and other previous approaches. OTSeg+ provides the most promising performance for both seen and unseen class objects compared to the previous SOTA methods. More visual results are provided in **??**.

[3] https://github.com/openai/CLIP

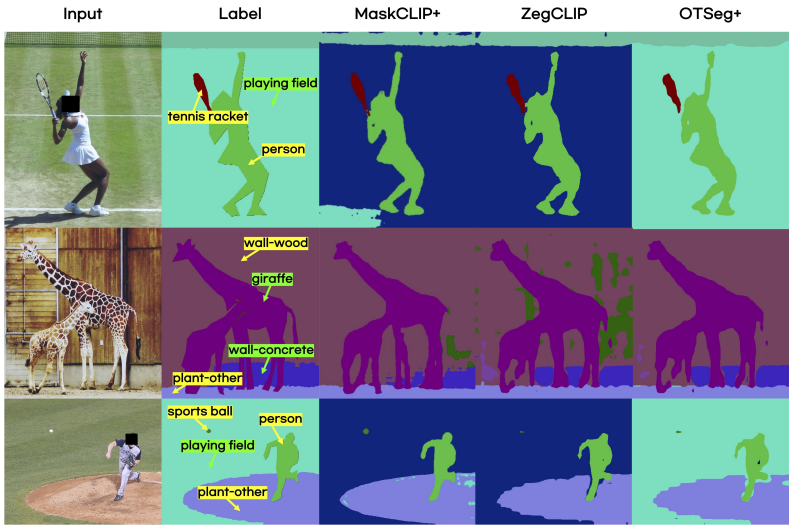

Fig. 4: Qualitative comparison with previous SOTA models on COCO-Stuff164K dataset. Green tag indicates unseen classes, while yellow indicates seen classes.

Table 2: Comparison in cross-data settings.

Method	Source	Target			Source	Target	
		ADE20K	PASCAL Context	VOC 2012		COCO-Stuff164K	VOC 2012
Inductive setting							
ZegFormer		16.4	-	80.7		-	-
Zsseg		15.3	-	74.5		-	-
ZegCLIP	COCO-156	19.0	41.2	93.4	Context-49	15.6	84.0
OTSeg		**20.5**	**49.3**	**94.1**		16.7	82.7
OTSeg+		19.6	49.1	**94.1**		**16.8**	**85.2**
Transductive setting							
ZegCLIP		21.1	45.8	94.2		18.1	90.6
OTSeg	COCO-156	**21.9**	52.9	94.2	Context-49	**18.9**	92.2
OTSeg+		21.1	**53.4**	**94.4**		18.1	**92.4**

Table 3: Comparison of memory cost and inference time. All models are evlauated on a single 3090 GPU.

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$_{-3.1}$

5.4 Comparison in Cross-Dataset Settings

To further evaluate the generalization capabilities of OTSeg, we perform cross-dataset experiments across COCO-Stuff164K and PASCAL Context datasets. The model is exclusively trained on the source dataset with labels for seen classes, as indicated as COCO-156 and Context-49, respectively, and then evaluated on the target dataset without any fine-tuning. As for the evaluation target, we added a challenging ADE20K [35] dataset. We compare the results in both inductive and transductive settings. As shown in Tab. 2, Our proposed methods, both OTSeg and OTSeg+, outperform the previous SOTA methods and demonstrate superior generalization performance in both experimental settings.

5.5 Comparison on Efficiency

To validate the efficiency of OTSeg, we compare the number of learnable parameters, training complexity (GFLOPS), and inference speed with other baseline

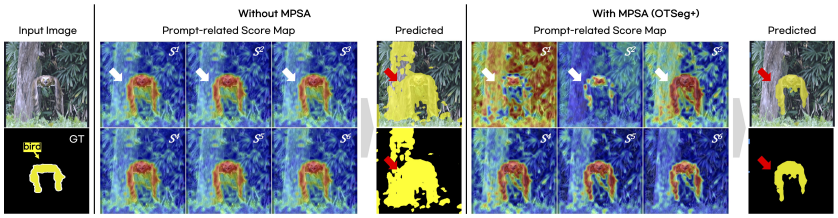

Fig. 5: Visual comparison of prompt-related score map. While all the text prompt-related score map S^i are cohered without MPSA, with our MPOT, each S^i is diversely activated and focuses on different semantic attributes (white arrows), which helps the model effectively differentiates the target object from the background (red arrows).

approaches in Tab. 3. Our OTSeg does not require additional learnable parameters compared to ZegCLIP, yet achieves the best performance while sacrificing slightly increased GFLOPS (1.2%) and decreased FPS (7.8%). Note that our method is still 4-5 times faster than other mask proposal-based two-stage methods such as Zsseg and ZegFormer.

5.6 Ablation Studies

MPSA for Enhanced Multimodal Alignment In Fig. 1, we visualize the effectiveness of MPSA by showing each text prompt-related score map, which focuses on different semantic features. We further analyze the reason behind the contribution of MPSA to produce better segmentation results by comparing each text prompt-related score map and their corresponding segmentation results in Fig. 5. With MPSA, we observe that each prompt-related score map is diversely dispersed, whereas the baseline method without MPSA shows significantly cohered score maps, as indicated in white arrows. This visual result suggests that MPSA is helpful in differentiating various semantic attributes from the target object, which yields the final score map selectively focuses on target-related features, leading improved performance on classnames-driven semantic segmentation.

Analysis on OTSeg Ensemble Component To investigate the effectiveness of our ensemble strategy which combines the decoder output and the refined score map, we ablate each path prediction from the ensembled prediction, as shown in Tab. 4. We provide the results for both the inductive and the transductive settings, as well as the fully supervised setting to establish the upper bound of each approach. Remarkably, depending solely on the decoder output or the refined score map, which corresponds to the effect of MPS or MPSA, demonstrates superior performance in all the setting. This suggests that the proposed MPS or MPSA effectively enhances zero-shot semantic segmentation performance. Despite each path prediction itself outperforms previous SOTA approaches, the proposed ensemble strategy yields the best performance in almost settings, validating the rationale of our ensemble approach.

Table 4: Analysis on OTSeg Ensemble Components.

Model Predictions		VOC 2012			PASCAL Context			COCO-Stuff164K		
Decoder	ScoreMap	mIoU(U)	mIoU(S)	hIoU	mIoU(U)	mIoU(S)	hIoU	mIoU(U)	mIoU(S)	hIoU
Inductive setting										
✓	✗	78.1	92.1	84.5	56.7	53.0	54.8	41.4	**41.4**	41.4
✗	✓	73.7	92.7	81.8	55.6	54.3	54.9	37.9	40.7	39.3
✓	✓	**81.6**	**93.3**	**87.1**	**60.4**	**55.2**	**57.7**	**41.8**	41.3	**41.5**
Transductive setting										
✓	✗	94.3	94.2	94.2	66.7	53.4	59.3	60.7	**41.8**	49.5
✗	✓	**94.4**	**94.3**	**94.3**	66.7	53.4	59.3	58.9	40.9	48.3
✓	✓	94.3	**94.3**	**94.3**	**67.0**	**54.0**	**59.8**	**62.6**	41.4	**49.8**
Fully-supervised										
✓	✗	94.4	94.0	94.2	78.1	55.2	64.7	**64.0**	**41.8**	**50.5**
✗	✓	94.7	**94.3**	94.5	77.8	55.4	64.7	62.5	40.9	49.5
✓	✓	**95.0**	94.1	**94.6**	**78.4**	**56.2**	**65.5**	63.2	41.5	50.1

Table 5: Component analysis of MPSA module under the inductive setting.

Component	Multi Text Prompt	Configuration	VOC 2012			PASCAL Context		
			mIoU(U)	mIoU(S)	hIoU	mIoU(U)	mIoU(S)	hIoU
(a) Number of Prompt	✓	4	76.5	93.3	84.1	44.8	52.0	48.1
		6	**81.6**	**93.3**	**87.1**	48.1	52.8	50.3
		8	79.9	93.3	86.1	**60.4**	**55.2**	**57.7**
(b) Matching Method	✗	Cross-attention	77.9	91.8	84.3	50.9	49.7	50.3
	✓		67.1	92.1	77.6	57.2	52.0	54.5
	✓	MPSA	**81.6**	**93.3**	**87.1**	**60.4**	**55.2**	**57.7**

5.7 Component Analysis

To study effect of the component of OTSeg on the zero-shot segmentation performance, we conduct a component analysis in Tab. 5, which includes: the number of text prompts and the matching method.

Number of Text Prompts In Tab. 5(a), we observe segmentation performance by varying the total number N of the introduced multiple text prompts. Our empirical findings indicate that OTSeg achieves optimal performance when $N = 6$ for VOC 2012 dataset, and $N = 8$ for PASCAL Context dataset. This suggests that, while $N = 6$ is sufficient for datasets with fewer classes, a way increased number of text prompts can add additional performance gains, particularly for larger datasets such like PASCAL Context and COCO-Stuff164K, which may assist the model in acquiring varied semantic features related to text prompts.

Matching Method In Tab. 5(b), we further compare our MPSA matching method with the conventional cross-attention and its variants. We observe that the naive extension of multiple text prompts for the cross-attention mechanism results in detrimental effects on specific datasets, such as VOC 2012. Whereas, our MPSA demonstrates its superior performance with margins of 3% and 7% hIoU compared to the cross-attention method. These results demonstrate the reason how OTSeg achieves the best performance in zero-shot segmentation settings, which is not only rooted from leveraging multiple text prompts, but also

from the proposed Sinkhorn matching mechanism, which optimally transports multiple text prompts to related pixel embeddings.

6 Discussion and Limitation

In this study, we demonstrate through MPS and MPSA that each text prompt serves as a score map capable of capturing different semantic features. However, our each score map is limited by the fact that the true meanings of text prompts are not fully captured. This issue arises from the framework that does not consider the association between semantic meanings and visual features. Furthermore, even though the application of zero-shot semantic segmentation yields noteworthy results with scalability, this module has yet to be applied to other tasks such as open-vocabulary or instance and panoptic segmentation within the scope of our investigation. These areas will be considered as future research.

7 Conclusion

In this study, we introduce OTSeg, a novel multimodal matching framework for zero-shot semantic segmentation, which leverages multiple text prompts with Optimal Transport (OT)-based text-pixel alignment module, specifically Multi-Prompts Sinkhorn (MPS), along with its extension for cross-attention mechanism, Multi-Prompts Sinkhorn Attention (MPSA). By incorporating MPSA within Transformer, our proposed OTSeg effectively aligns semantic features between multiple text prompts and image pixels, and selectively focuses on target object-related features. Through extensive experiments, we demonstrate OTSeg's capability to achieve the state-of-the-art (SOTA) performance on zero-shot segmentation (ZS3) tasks across three benchmark datasets. We believe that OTSeg can contribute in opening new directions for future researches in multimodal alignment and zero-shot learning, with potential applications in various domains requiring multi-conceptual semantic understandings of vision.

Acknowledgments This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.RS-2019-II190075) Artificial Intelligence Graduate School Program(KAIST), and also supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2024-00345854), (RS-2024-00336454), (RS-2023-00262527), and also supported by Field-oriented Technology Development Project for Customs Administration funded by the Korea government through the National Research Foundation (NRF) of Korea under Grant NRF2021M3I1A1097910, and also Administration by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2023.

References

1. Allingham, J.U., Ren, J., Dusenberry, M.W., Gu, X., Cui, Y., Tran, D., Liu, J.Z., Lakshminarayanan, B.: A simple zero-shot prompt weighting technique to improve prompt ensembling in text-image models. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 547–568. PMLR (23–29 Jul 2023), https://proceedings.mlr.press/v202/allingham23a.html 2
2. Bucher, M., Vu, T.H., Cord, M., Pérez, P.: Zero-shot semantic segmentation. Advances in Neural Information Processing Systems **32** (2019) 2, 3, 9
3. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1209–1218 (2018) 9
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9650–9660 (2021) 1
5. Chen, G., Yao, W., Song, X., Li, X., Rao, Y., Zhang, K.: Prompt learning with optimal transport for vision-language models. arXiv preprint arXiv:2210.01253 (2022) 3
6. Chen, L., Gan, Z., Cheng, Y., Li, L., Carin, L., Liu, J.: Graph optimal transport for cross-domain alignment. In: International Conference on Machine Learning. pp. 1542–1553. PMLR (2020) 3
7. Contributors, M.: MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation (2020) 10
8. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26** (2013) 3, 4
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018) 1
10. Ding, J., Xue, N., Xia, G.S., Dai, D.: Decoupling zero-shot semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11583–11592 (2022) 2, 3, 9, 10
11. Everingham, M., Winn, J.: The pascal visual object classes challenge 2012 (voc2012) development kit. Pattern Anal. Stat. Model. Comput. Learn., Tech. Rep **2007**, 1–45 (2012) 9
12. Flamary, R., Courty, N., Tuia, D., Rakotomamonjy, A.: Optimal transport for domain adaptation. IEEE Trans. Pattern Anal. Mach. Intell **1** (2016) 3
13. Gu, Z., Zhou, S., Niu, L., Zhao, Z., Zhang, L.: Context-aware feature generation for zero-shot semantic segmentation. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 1921–1929 (2020) 3
14. Guo, J., Wang, Q., Gao, Y., Jiang, X., Tang, X., Hu, Y., Zhang, B.: Mvp-seg: Multi-view prompt learning for open-vocabulary semantic segmentation. arXiv preprint arXiv:2304.06957 (2023) 3, 10
15. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European Conference on Computer Vision. pp. 709–727. Springer (2022) 10
16. Kim, J., El-Khamy, M., Lee, J.: T-gsa: Transformer with gaussian-weighted self-attention for speech enhancement. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6649–6653. IEEE (2020) 1

17. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. arXiv preprint arXiv:2201.03546 (2022) 3
18. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. arXiv preprint arXiv:2210.04150 (2022) 3
19. Liu, Y., Zhu, L., Yamada, M., Yang, Y.: Semantic correspondence as an optimal transport problem. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4463–4472 (2020) 3
20. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 891–898 (2014) 9
21. Nguyen, T.M., Nguyen, T.M., Le, D.D., Nguyen, D.K., Tran, V.A., Baraniuk, R., Ho, N., Osher, S.: Improving transformers with probabilistic attention keys. In: International Conference on Machine Learning. pp. 16595–16621. PMLR (2022) 1
22. Nguyen, T., Nguyen, T., Do, H., Nguyen, K., Saragadam, V., Pham, M., Nguyen, K.D., Ho, N., Osher, S.: Improving transformer with an admixture of attention heads. Advances in neural information processing systems **35**, 27937–27952 (2022) 1
23. Pastore, G., Cermelli, F., Xian, Y., Mancini, M., Akata, Z., Caputo, B.: A closer look at self-training for zero-label semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2693–2702 (2021) 3
24. Peyré, G., Cuturi, M., et al.: Computational optimal transport: With applications to data science. Foundations and Trends® in Machine Learning **11**(5-6), 355–607 (2019) 3
25. Qin, J., Wu, J., Yan, P., Li, M., Yuxi, R., Xiao, X., Wang, Y., Wang, R., Wen, S., Pan, X., et al.: Freeseg: Unified, universal and open-vocabulary image segmentation. arXiv preprint arXiv:2303.17225 (2023) 3, 10
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021) 1, 2, 3
27. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Denseclip: Language-guided dense prediction with context-aware prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18082–18091 (2022) 2
28. Sander, M.E., Ablin, P., Blondel, M., Peyré, G.: Sinkformers: Transformers with doubly stochastic attention. In: International Conference on Artificial Intelligence and Statistics. pp. 3515–3530. PMLR (2022) 1, 4
29. Schmitzer, B.: Stabilized sparse scaling algorithms for entropy regularized transport problems. SIAM Journal on Scientific Computing **41**(3), A1443–A1481 (2019) 5
30. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023) 1
31. Xu, H., Luo, D., Carin, L.: Scalable gromov-wasserstein learning for graph partitioning and matching. Advances in neural information processing systems **32** (2019) 3
32. Xu, H., Luo, D., Zha, H., Duke, L.C.: Gromov-wasserstein learning for graph matching and node embedding. In: International conference on machine learning. pp. 6932–6941. PMLR (2019) 3

33. Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X.: Groupvit: Semantic segmentation emerges from text supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18134–18144 (2022) 3
34. Xu, M., Zhang, Z., Wei, F., Lin, Y., Cao, Y., Hu, H., Bai, X.: A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model. arXiv preprint arXiv:2112.14757 (2021) 2, 3, 9, 10
35. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. International Journal of Computer Vision **127**, 302–321 (2019) 11
36. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: European Conference on Computer Vision. pp. 696–712. Springer (2022) 2, 3, 9, 10
37. Zhou, Z., Lei, Y., Zhang, B., Liu, L., Liu, Y.: Zegclip: Towards adapting clip for zero-shot semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11175–11185 (2023) 2, 3, 7, 9, 10