EAGLE: Eigen Aggregation Learning for Object-Centric Unsupervised Semantic Segmentation

Chanyoung Kim* Woojung Han* Dayun Ju Seong Jae Hwang†
Yonsei University
{chanyoung, dnwjddl, juda0707, seongjae}@yonsei.ac.kr

Abstract

*Semantic segmentation has innately relied on extensive pixel-level annotated data, leading to the emergence of unsupervised methodologies. Among them, leveraging self-supervised Vision Transformers for unsupervised semantic segmentation (USS) has been making steady progress with expressive deep features. Yet, for semantically segmenting images with complex objects, a predominant challenge remains: the lack of explicit object-level semantic encoding in patch-level features. This technical limitation often leads to inadequate segmentation of complex objects with diverse structures. To address this gap, we present a novel approach, **EAGLE**, which emphasizes object-centric representation learning for unsupervised semantic segmentation. Specifically, we introduce EiCue, a spectral technique providing semantic and structural cues through an eigenbasis derived from the semantic similarity matrix of deep image features and color affinity from an image. Further, by incorporating our object-centric contrastive loss with EiCue, we guide our model to learn object-level representations with intra- and inter-image object-feature consistency, thereby enhancing semantic accuracy. Extensive experiments on COCO-Stuff, Cityscapes, and Potsdam-3 datasets demonstrate the state-of-the-art USS results of EAGLE with accurate and consistent semantic segmentation across complex scenes.*

1. Introduction

Semantic segmentation plays a pivotal role in modern vision, fundamentally advancing an array of diverse areas including medical imaging [21, 40], autonomous driving [14, 46], and remote sensing imagery [12, 28]. Nevertheless, its reliance on labeled data, while common across nearly all vision tasks, is especially problematic due to the laborious and time-consuming process of pixel-level annotation. In response to this challenge, various studies in semantic segmentation tasks have drifted away from relying solely on human-labeled annotations by exploring weakly-supervised [1, 23, 26, 39, 48], semi-supervised [2, 27, 37], and *unsupervised semantic segmentation* (USS) methodologies [8, 15, 16, 20, 22, 36, 43, 52].

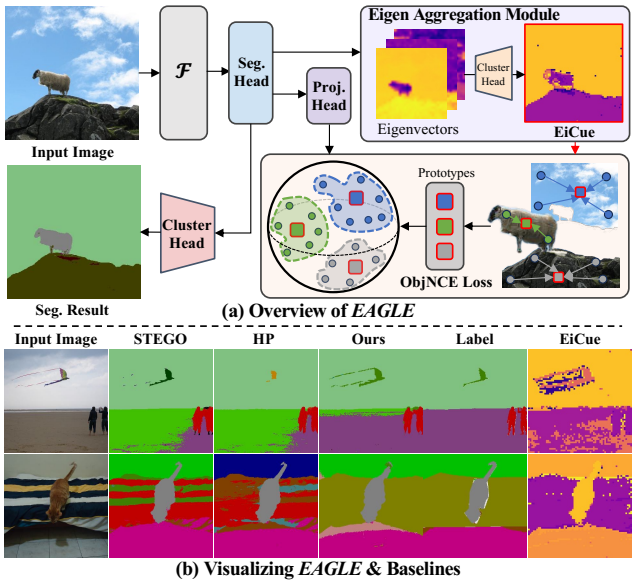

Figure 1. We introduce ***EAGLE***, ***E**igen **AG**gregation **LE**arning* for object-centric unsupervised semantic segmentation. **(a)** We first leverage the aggregated eigenvectors, named EiCue, to obtain the semantic structure knowledge of object segments in an image. Based on both semantic and structural cues from the EiCue, we compute object-centric contrastive loss to learn object-level semantic representation. **(b)** A visual comparison between *EAGLE* and other methods. Our object-level semantic segmentation results robustly identify objects with complex semantics (e.g., `blanket` with vivid stripe patterns) by exploiting strong semantic structure cues from EiCue.

Among these learning schemes, the unsupervised approach of USS clearly stands as the most challenging case. Specifically, compared to the classical unsupervised segmentation methods (e.g., K-means clustering) which produce segments without explicit semantics, USS additionally

*Equal contribution
†Corresponding author
Project Page: https://micv-yonsei.github.io/eagle2024/

aims to derive semantically consistent local features (e.g., patch-level features) that aid the further class assignment post-steps via clustering and the Hungarian matching algorithm. That is, semantically plausible local features result in accurate semantic segmentation results (e.g., Fig. 1b), but in USS, this must be achieved *without any labels*.

Despite the glaring challenge, steady progress has been shown in USS. For example, initial pioneering works have emerged to maximize the mutual information across the two different views of a single image [20, 36]. Recently, network-based techniques such as STEGO [15] have focused on deriving patch-level semantic features with a self-supervised pretrained model [6], showing a significant improvement compared to previous methods [8, 20, 52]. However, while these methodologies have advanced USS, unresolved shortcomings still remain.

In particular, the recent network-based methods often leverage a self-supervised Vision Transformer (ViT) to learn patch-level features. While their patch-level features proved to be useful for further USS inference steps (e.g., K-means), the underlying object-level semantics are not explicitly imposed in these patch-level features. To grasp the "object-level semantics", consider an example of a `blanket` object as shown in Fig. 1b second row. As with any object, `blanket` may easily appear with varying colors and textures across different images. Without proper object-level semantics, features corresponding to varying regions of `blanket` may result in vastly different feature representations. Ideally, though, the features corresponding to all kinds of `blanket` should be mapped to similar features, namely, object-level semantics. Thus, without carefully imposed object-level semantics, complex objects with diverse structures and shapes may easily be partitioned into multiple segments with wrong class labels or be merged with nearby segments of different class labels. Thus, in USS, an immense effort must be paid to learn the local features (e.g., patch-level) with strong object-level semantics.

Our object-centric representation learning for USS aims to capture such object-level semantics. Specifically, we first need a semantic or structure cue in the object-centric view. Several previous works utilized clustering methods such as K-means or superpixel to obtain semantic cues [19], however, they mainly fixated on the generic image patterns, not the object's semantic or structural representation. Here, we propose EiCue which provides semantic and structural cues of objects via eigenbasis. Specifically, we utilize the semantic similarity matrix obtained from the projected deep image features obtained from ViT [6, 13] and the color affinity matrix of the image to construct the graph Laplacian. The corresponding eigenbasis captures the underlying *semantic structures* of objects [29, 55], providing soft guidance to the subsequent object-level feature refinement step.

Recall that accurate object-level semantics of an object must be consistent across images. Our object-centric contrastive learning framework explicitly imposes these traits with a novel object-level contrastive loss. Specifically, based on the object cues from EiCue, we derive learnable prototypes for each object which enables intra- and inter-image object-feature consistency. Through this comprehensive learning process, our model effectively captures the inherent structures within images, allowing it to precisely identify semantically plausible object representations, the key to advancing modern feature-based USS.

Contributions. Our main contributions are as follows:

- We propose EiCue, using a learnable graph Laplacian, to acquire a more profound understanding of the underlying semantics and structural details within images.
- We design an object-centric contrastive learning framework that capitalizes on the spectral basis of EiCue to construct robust object-level feature representations.
- We demonstrate that our *EAGLE* achieves state-of-the-art performance on unsupervised semantic segmentation, supported by a series of comprehensive experiments.

2. Related Work

2.1. Unsupervised Semantic Segmentation

Semantic segmentation plays a crucial role in vision by assigning distinct class labels to pixels. Yet, while the segmentation performance strongly correlates with the label quality, acquiring precise pixel-level ground truth labels is a challenge on its own, especially for images with complex structures. This naturally led to numerous attempts to perform semantic segmentation in an unsupervised manner [8, 15, 16, 20, 22, 36, 43, 52], that is, with no labels. For instance, early works such as IIC [20] and AC [36] utilized mutual information, while subsequent approaches like InfoSeg [16] and PiCIE [8] integrated diverse features for enhanced pixel learning. Recent studies have adopted self-supervised, pretrained ViT models like DINO [6] for top-down feature extraction. Namely, STEGO [15] demonstrated a major step forward by distilling unsupervised features into discrete semantic labels with the DINO backbone. HP [43] interestingly utilizes contrastive learning to enhance semantic correlations among patch-level regions, but this patch-level (local) refinement holds little object-level understanding.

2.2. Spectral Techniques for Segmentation

Predating the aforementioned methods for semantic segmentation, spectral techniques have long been offering insights into diverse segmentation challenges in vision. Spanning some early pioneering works [30, 34, 38, 45] to contemporary efforts [3, 11, 24, 32, 44], these techniques share a common aim: to exploit the intrinsic spectral signatures embedded within image regions. These graph-theoretic approaches are methodologically influenced by the affinity

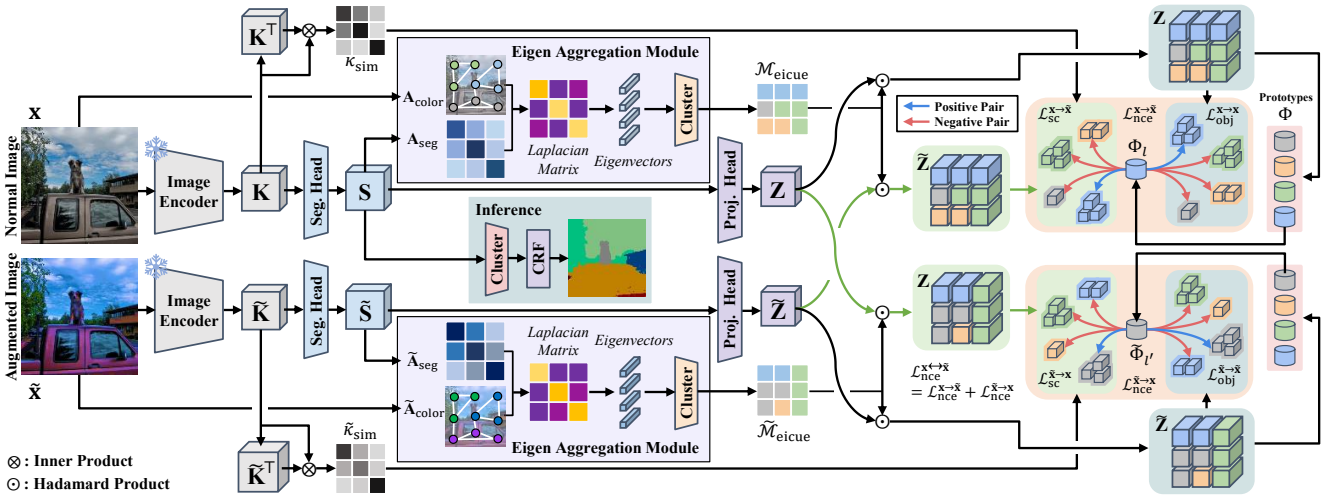

Figure 2. The pipeline of ***EAGLE***. Leveraging the Laplacian matrix, which integrates hierarchically projected image key features and color affinity, the model exploits eigenvector clustering to capture object-level perspective cues defined as $\mathcal{M}_{\text{eicue}}$ and $\tilde{\mathcal{M}}_{\text{eicue}}$. Distilling knowledge from $\mathcal{M}_{\text{eicue}}$, our model further adopts an object-centric contrastive loss, utilizing the projected feature $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$. The learnable prototype Φ assigned from $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$, acts as a singular anchor that contrasts positive objects and negative objects. Our object-centric contrastive loss is computed in two distinct manners: intra($\mathcal{L}_{\text{obj}}$)- and inter($\mathcal{L}_{\text{sc}}$)-image to ensure semantic consistency.

matrix quality, which gave rise to recent methods utilizing the network features from the pretrained deep models. For instance, Deep Spectral Methods [33] builds powerful Laplacian eigenvectors from the feature affinity matrix, while EigenFunction [10] exploits the network-based learnable eigenfunctions to produce spectral embeddings. Despite steadily discovering the effectiveness of spectral methods on deep features for capturing complex object structures, their object-level semantics still require additional methodological efforts, e.g., contrastive learning.

2.3. Object-centric Contrastive Learning

Contrastive learning approaches aim to maximize feature similarities between similar units while minimizing them between dissimilar ones. In the task of semantic segmentation, patch-level representation learning [35, 49, 51] is widely used. However, this approach tends to overemphasize fine details while neglecting high-level concepts (i.e., semantic relations between objects). This leads object-level contrastive learning methods [17, 41, 42, 47, 50, 53, 54] to focus on balancing detailed perception with an object-centric view, identifying objects in an unsupervised manner. For instance, MaskContrast [47] and COMUS [53] use unsupervised saliency to make pixel embeddings, while Odin [18] and DetCon [17] utilize K-means clustering and heuristic masks for sample generation, respectively. Refining this, SlotCon [50] assigned pixels to learn slots for semantic representation, and DINOSAUR [41] further improved it by reconstructing self-supervised pretrained features in the decoder, instead of the original inputs. However, these methods [41, 50] rely solely on slots, potentially overlooking high-level image features. In contrast, our approach distills knowledge from clustered eigenvectors derived from a similarity matrix-based Laplacian capturing their object semantic relationships.

3. Methods

As we begin describing our full pipeline shown in Fig. 2, let us first cover the core USS framework based on pretrained models as in prior works [15, 43].

3.1. Preliminary

Unlabeled Images. Our approach is built exclusively upon a set of images, *without* any annotations, denoted as $\mathbf{X} = \{\mathbf{x}_b\}_{b=1}^{B}$, where B is the number of training images within a mini-batch. We also utilize a photometric augmentation strategy P to obtain an augmented image set $\tilde{\mathbf{X}} = \{\tilde{\mathbf{x}}_b\}_{b=1}^{B} = P(\mathbf{X})$.

Pretrained Features K. Then, for each input image $\mathbf{x}_b$, we use a self-supervised pretrained vision transformer [6] as an image encoder $\mathcal{F}$ to obtain hierarchical attention key features from the last three blocks as $\mathbf{K}_{L-2} = \mathcal{F}_{L-2}(\mathbf{x}_b)$, $\mathbf{K}_{L-1} = \mathcal{F}_{L-1}(\mathbf{x}_b)$, $\mathbf{K}_{L} = \mathcal{F}_{L}(\mathbf{x}_b)$, where $L-2$, $L-1$, L is the third-to-last layer, the second-to-last layer, and the last layer, respectively. Then, we concatenate them into a single attention tensor $\mathbf{K} = [\mathbf{K}_{L-2}; \mathbf{K}_{L-1}; \mathbf{K}_{L}] \in \mathbb{R}^{H \times W \times D_K}$. Similarly, we apply the same procedure for the augmented image $\tilde{\mathbf{x}}_b$ and obtain its attention tensor $\tilde{\mathbf{K}} \in \mathbb{R}^{H \times W \times D_K}$.

Semantic Features S. Although $\mathbf{K}$ contains some structural information about the objects based on the attention mechanism, this is known for insufficient semantic information to be considered for direct inference. Thus, for further feature refinement, we compute the semantic features $\mathbf{S} = \mathcal{S}_\theta(\mathbf{K}) \in \mathbb{R}^{H \times W \times D_S}$ and $\tilde{\mathbf{S}} = \mathcal{S}_\theta(\tilde{\mathbf{K}}) \in \mathbb{R}^{H \times W \times D_S}$, where $\mathcal{S}_\theta : \mathbb{R}^{H \times W \times D_K} \rightarrow \mathbb{R}^{H \times W \times D_S}$ is a learnable nonlinear segmentation head. For brevity, the total number of

patches, denoted as $H \times W$, will be referred to as N.

Inference. During the inference time, given a new image, its semantic feature $\mathbf{S}$ becomes the basis of further clustering for the final semantic segmentation output with conventional evaluation setups such as the K-means clustering and linear probing. Thus, as with prior pretrained feature-based USS works [15, 43], training $\mathcal{S}_\theta$ to output strong semantic features $\mathbf{S}$ in an unsupervised manner is the basic framework of contemporary USS frameworks. We next describe the remainder of the pipeline in Fig. 2 which corresponds to our methodological contributions for producing powerful *object-level* semantic features.

3.2. EiCue via the Eigen Aggregation Module

Intuition tells us that the "semantically plausible" object-level segments are groups of pixels precisely capturing the object structure, even under complex structural variance. For instance, a `car` segment must contain all of its parts including the windshield, doors, wheels, etc. which may all appear in different shapes and views. However, without pixel-level annotations that provide object-level semantics, this becomes an extremely challenging task of inferring the underlying structure with zero object-level structural prior.

From this realization, our model *EAGLE* first aims to derive a strong yet simple semantic structural cue, namely, EiCue, based on the eigenbasis of the feature similarity matrix as illustrated in Fig. 3. Specifically, we use the well-known Spectral Clustering [7, 34, 45] to obtain unsupervised feature representations that capture the underlying non-linear structures for handling data with complex patterns. This classically operates only in the color space but may easily extend to utilize the similarity matrix constructed from any features. We observed that such a spectral method becomes especially useful for complex real-world images as in Fig. 4.

EiCue Construction. Let us describe the process of constructing EiCue in detail as shown in Fig. 3. The overall framework generally follows the vanilla spectral clustering: (1) from an adjacency matrix $\mathbf{A}$, (2) construct the graph Laplacian $\mathbf{L}$, and (3) perform the eigendecomposition on $\mathbf{L}$ to derive the eigenbasis $\mathbf{V}$ from which the eigenfeatures are used for the clustering. We describe each step below.

3.2.1 Adjacency Matrix Construction

Our adjacency matrix consists of two components: (1) color affinity matrix and (2) semantic similarity matrix.

(I) Color Affinity Matrix $\mathbf{A}_{\text{color}}$: The color affinity matrix leverages the RGB values of the image $\mathbf{x}$. The color affinity matrix is computed by the color distance. It utilizes the Euclidean distance between patches, where p and q are specific patch positions within the image. Here, $\ddot{\mathbf{x}} \in \mathbb{R}^{H \times W \times 3}$ denotes a resized version of $\mathbf{x}$, scaled from its original image resolution to patch resolution, to ensure compatibility with the dimensions of other adjacency matrices. The resulting

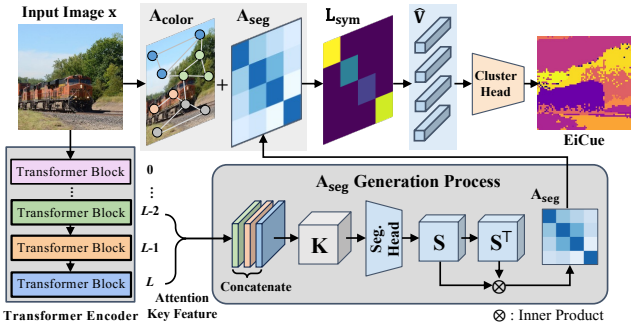

Figure 3. An illustration of the EiCue generation process. From the input image, both color affinity matrix $\mathbf{A}_{\text{color}}$ and semantic similarity matrix $\mathbf{A}_{\text{seg}}$ are derived, which are combined to form the Laplacian $\mathbf{L}_{\text{sym}}$. An eigenvector subset $\hat{\mathbf{V}}$ of $\mathbf{L}_{\text{sym}}$ are clustered to produce EiCue.

color affinity matrix, $\mathbf{A}_{\text{color}} \in \mathbb{R}^{N \times N}$ thus captures the pairwise relationship between the patches based on the colors. Specifically, we use the RBF kernel as the distance function $\mathbf{A}_{\text{color}}(p, q) = \exp\left(-\|\ddot{\mathbf{x}}(p) - \ddot{\mathbf{x}}(q)\|_2 / 2{\sigma_c}^2\right)$ where $\sigma_c > 0$ is a free hyperparameter. Further, to ensure that only nearby patches influence each other's affinity values, we hard-constrain the maximum distance of the patch pairs such that we only compute the affinity between the patch pairs with a predefined spatial distance.

(II) Semantic Similarity Matrix $\mathbf{A}_{\text{seg}}$: The semantic similarity matrix, denoted as $\mathbf{A}_{\text{seg}} \in \mathbb{R}^{N \times N}$, is formed by the product of tensor $\mathbf{S}$ and its transpose $\mathbf{S}^\top$. Tensor $\mathbf{S}$ is derived by hierarchically concatenating key attention features from the last three layers of a pretrained vision transformer, as processed through the segmentation head $\mathcal{S}_\theta$.

(III) Adjacency Matrix $\mathbf{A}$: The final adjacency matrix $\mathbf{A}$ is the sum of $\mathbf{A}_{\text{color}}$ and $\mathbf{A}_{\text{seg}}$: $\mathbf{A} = \mathbf{A}_{\text{color}} + \mathbf{A}_{\text{seg}}$, which is also applicable to $\tilde{\mathbf{A}}$. Our adjacency matrix amalgamates the high-level color information and the network-based deep features to characterize semantic-wise relations. The use of the image-based $\mathbf{A}_{\text{color}}$ preserves the image's structural integrity and also complements the contextual information of the image. Following this, the incorporation of the learnable tensor $\mathbf{S}$ for the $\mathbf{A}_{\text{seg}}$ further strengthens this aspect, enhancing the semantic interpretation of the object without compromising the structural integrity and serving as a vital cue for our learning process.

3.2.2 Eigendecomposition

To construct EiCue based on $\mathbf{A}$, a Laplacian matrix is created. Formally, the Laplacian Matrix is expressed as $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where $\mathbf{D}$ is the degree matrix of $\mathbf{A}$ defined as $\mathbf{D}(i, i) = \sum_{j=1}^{N} \mathbf{A}(i, j)$. In our approach, we utilize the normalized Laplacian matrix for its enhanced clustering capabilities. The symmetric normalized Laplacian matrix $\mathbf{L}_{\text{sym}}$ are defined as $\mathbf{L}_{\text{sym}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}}$. Then, via eigendecomposition on $\mathbf{L}_{\text{sym}}$, the eigenbasis $\mathbf{V} \in \mathbb{R}^{N \times N}$ is computed, where each column corresponds to a unique

eigenvector. We then extract k eigenvectors corresponding to the k smallest eigenvalues and concatenate them into $\hat{\mathbf{V}} \in \mathbb{R}^{N\times k}$ where the i^{th} row corresponds to the k dimensional eigenfeature of the i^{th} patch.

3.2.3 Differentiable Eigen Clustering

After obtaining eigenvectors $\hat{\mathbf{V}}$, we perform the eigenvectors clustering process and extract the EiCue denoted as $\mathcal{M}_{\text{eicue}} \in \mathbb{R}^N$. To cluster eigenvectors, we leverage a mini-batch K-means algorithm based on cosine distance [31] between $\hat{\mathbf{V}}$ and $\mathbf{C}$, denoted as $\mathbf{P} = \hat{\mathbf{V}}\mathbf{C}$. Centers of clusters $\mathbf{C} \in \mathbb{R}^{k\times C}$ are composed of learnable parameters. To learn $\mathbf{C}$, we further trained with a loss defined as follows:

$$\mathcal{L}_{\text{eig}}^{\mathbf{x}} = -\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{c=1}^{C}\Psi_{ic}\mathbf{P}_{ic}\right), \tag{1}$$

where C denotes pre-defined number of classes, $\Psi := \textit{softmax}(\mathbf{P})$ and $\mathbf{P}_{ic}$ and Ψ_{ic} represents the i^{th} patch and the c^{th} cluster number of $\mathbf{P}$ and Ψ. We apply same procedure to augmented image $\tilde{\mathbf{x}}$ to get $\mathcal{L}_{\text{eig}}^{\tilde{\mathbf{x}}}$. By minimizing $\mathcal{L}_{\text{eig}} = \frac{1}{2}(\mathcal{L}_{\text{eig}}^{\mathbf{x}} + \mathcal{L}_{\text{eig}}^{\tilde{\mathbf{x}}})$, we can obtain centers of clusters that enable more effective clustering. Then we obtain EiCue as

$$\mathcal{M}_{\text{eicue}}(i) = \underset{c}{\text{argmax}}\left(\mathbf{P}_{ic} - \log\left(\sum_{c'=1}^{C}\exp(\mathbf{P}_{ic'})\right)\right). \tag{2}$$

As the precision of cluster centroids improves, EiCue facilitates the mapping of patch i to its corresponding object based on semantic structure. This serves as a meaningful cue to stress semantic distinctions between different objects, thereby enhancing the discriminative power of the feature embeddings.

Remark. While similar to previous work [33] in using eigendecomposition, our approach differs by enhancing feature vectors $\mathbf{S}$ with a trainable segmentation head, unlike their reliance on static vectors (i.e., $\mathbf{K}$). Our method enhances $\mathbf{S}$ learnable and adaptable via differentiable eigen clustering, allowing the graph Laplacian and object semantics to evolve. This dynamic integration of EiCue into the learning process distinctly separates our methodology from prior applications.

3.3. EiCue-based ObjNCELoss

For a successful semantic segmentation task, it is important not only to classify the class of each pixel accurately but also to aggregate object representation and create a segmentation map that reflects object semantic representations. From this perspective, learning relationships in an object-centric view is especially crucial in semantic segmentation tasks. To capture the complex relationships between objects, our approach incorporates an object-centric contrastive learning strategy, named ***ObjNCELoss***, guided by EiCue. This strategy is designed to refine the discriminative capabilities of feature embeddings $\mathbf{S}$, emphasizing the

Figure 4. Visualizing eigenvectors derived from $\mathbf{S}$ in the Eigen Aggregation Module. These eigenvectors not only distinguish different objects but also identify semantically related areas, highlighting how EiCue captures object semantics and boundaries effectively.

distinctions among various object semantics. Before proceeding, we map both the projected feature $\mathbf{Z} \in \mathbb{R}^{N\times D_Z}$ and $\tilde{\mathbf{Z}} \in \mathbb{R}^{N\times D_Z}$, using the linear projection head $\mathcal{Z}_\xi$, derived from the reshaped $\mathbf{S} \in \mathbb{R}^{N\times D_S}$ and $\tilde{\mathbf{S}} \in \mathbb{R}^{N\times D_S}$, respectively. While the actual dimension sizes of D_S and D_Z are kept the same, we use different notations for ease of explanation.

3.3.1 Object-wise Prototypes

To extract the representative object level semantic features from projected feature $\mathbf{Z}$, we construct adaptable prototypes Φ_l based on the object l in aforementioned EiCue. As we describe next, semantically representative prototypes become the anchors for either pulling objects with similar semantics while pushing away the different ones.

Let us describe how Φ is derived, which represents object semantics from $\mathbf{Z}$. We first update the object-wise prototypes through the projected feature $\mathbf{Z}$ and a given $\mathcal{M}_{\text{eicue}}$, derived from the clustered eigenbasis. Formally, for each object l obtained from $\mathcal{M}_{\text{eicue}}$, the mask M_l is defined as $M_l(i) = 1$ if $\mathcal{M}_{\text{eicue}}(i) = l$, and 0 otherwise, where i represents each position in $\mathcal{M}_{\text{eicue}}$. Then, applying the mask M_l to the projected feature tensor $\mathbf{Z}$ gives $\mathbf{Z}_l = \mathbf{Z} \odot M_l$, where $\odot$ denotes the Hadamard product and $\mathbf{Z}_l$ represents a collection of feature representations from $\mathbf{Z}$ corresponding to object l. Next, we compute medoid to select a single vector from $\mathbf{Z}_l$, which then becomes the prototype Φ_l. Let $\mathcal{I}_l$ be the set of indices where $M_l^{(i\in\mathcal{I}_l)} = 1$ to only consider the indices of object l. $\mathbf{Z}_l^{(i)}$ indicates the i-th feature vector of $\mathbf{Z}_l$. Then, the prototype Φ_l from the masked tensor $\mathbf{Z}_l$ is

$$\Phi_l = \mathbf{Z}_l^{(m^*)} \text{ for } m^* = \underset{m\in\mathcal{I}_l}{\text{argmin}}\sum_{i\in\mathcal{I}_l}\left\|\mathbf{Z}_l^{(m)} - \mathbf{Z}_l^{(i)}\right\|_2. \tag{3}$$

Thus, Φ_l acts as the semantic vector of object l, serving as an anchor for the following object-centric contrastive loss.

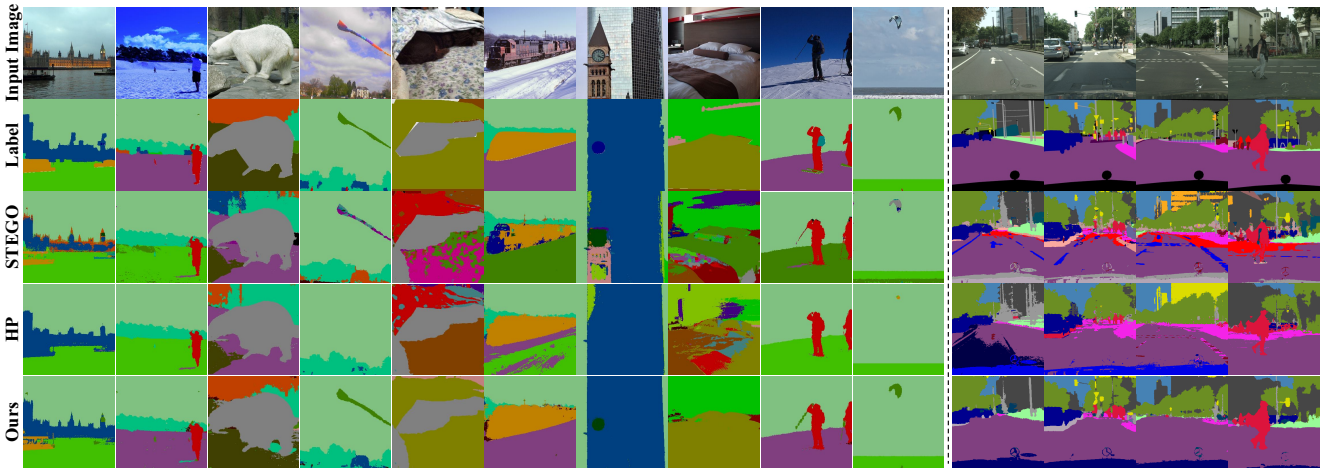

Figure 5. A qualitative comparison of the (a) COCO-Stuff [4] and (b) Cityscapes [9] datasets trained using ViT-S/8 and ViT-B/8 as a backbone, respectively. The comparison included previous state-of-the-art USS approaches, STEGO [15], HP [43], and ours.

3.3.2 Object-centric Contrastive Loss

Once we compute prototypes, we then step towards object-centric contrastive loss between prototypes Φ and feature vectors $\mathbf{Z}$. Specifically, we compute object-centric contrastive loss defined as follows:

$$\mathcal{L}_{\text{obj}}^{\mathbf{x}\to\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} w_{\text{obj}}^{(i)} \left[-\log\left(\frac{\exp((\mathbf{Z}_l^{(i)}\cdot\Phi_l)/\tau)}{\sum_{j=1,j\neq l}^{C}\exp((\mathbf{Z}_j\cdot\Phi_l)/\tau)} \right) \right], \tag{4}$$

where C is the total number of unique predicted objects in $\mathcal{M}_{\text{eicue}}$, $(\cdot)$ denotes the cosine similarity, and $\tau > 0$ is the temperature scalar. To emphasize the influence of feature vectors with high similarity and direct the model's focus toward them, we weigh the loss based on the similarity information between vectors. The weight $w_{\text{obj}}^{(i)}$ is defined as $w_{\text{obj}}^{(i)} = (\sum_{j=1}^{N}\mathcal{K}_{\text{sim}}(i,j))/N$, where $\mathcal{K}_{\text{sim}} \in \mathbb{R}^{N\times N}$ represents the similarity matrix defined as $\mathcal{K}_{\text{sim}} = \mathbf{K}\mathbf{K}^{\top}$.

While Eq. (4) aggregates the object-level features based on the EiCue assignment, we note that another kind of robust consistency could be cleverly imposed with our photometric augmented image $\tilde{\mathbf{x}}$. That is, since the photometric augmentation does not apply structural changes, the augmented image $\tilde{\mathbf{x}}$ and $\mathbf{x}$ are structurally identical, allowing us to make the following important assumption: the vectors in the same positions of $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ should have similar object-level semantics. This assumption ultimately allows us to create a new masked $\tilde{\mathbf{Z}}$ (Fig. 2, $\tilde{\mathbf{Z}}$ in green box) of $\tilde{\mathbf{x}}$ based on $\mathcal{M}_{\text{eicue}}$ of $\mathbf{x}$. Thus, we apply the contrastive loss to the augmented image $\tilde{\mathbf{x}}$, based on the prototypes Φ from the non-augmented image $\mathbf{x}$ to guide the model to learn global semantic consistency. To illustrate this concept, our semantic consistency contrastive loss is defined as

$$\mathcal{L}_{\text{sc}}^{\mathbf{x}\to\tilde{\mathbf{x}}} = \frac{1}{N}\sum_{i=1}^{N} w_{\text{obj}}^{(i)} \left[-\log\left(\frac{\exp((\tilde{\mathbf{Z}}_l^{(i)}\cdot\Phi_l)/\tau)}{\sum_{j=1,j\neq l}^{C}\exp((\tilde{\mathbf{Z}}_j\cdot\Phi_l)/\tau)} \right) \right], \tag{5}$$

where $\tilde{\mathbf{Z}}_l^{(i)}$ notes the i-th feature vector of projected feature $\tilde{\mathbf{Z}}$ for object l. Concretely, we can formulate our object-centric contrastive loss as $\mathcal{L}_{\text{nce}}^{\mathbf{x}\to\tilde{\mathbf{x}}} = \lambda_{\text{obj}}\mathcal{L}_{\text{obj}}^{\mathbf{x}\to\mathbf{x}} + \lambda_{\text{sc}}\mathcal{L}_{\text{sc}}^{\mathbf{x}\to\tilde{\mathbf{x}}}$, where $0 < \lambda_{\text{obj}} < 1$ and $0 < \lambda_{\text{sc}} < 1$ are hyperparameters that adjust the strength of each loss. Since the loss function $\mathcal{L}_{\text{nce}}^{\mathbf{x}\to\tilde{\mathbf{x}}}$ is asymmetric, we also take into account the opposite case as $\mathcal{L}_{\text{nce}}^{\tilde{\mathbf{x}}\to\mathbf{x}} = \lambda_{\text{obj}}\mathcal{L}_{\text{obj}}^{\tilde{\mathbf{x}}\to\tilde{\mathbf{x}}} + \lambda_{\text{sc}}\mathcal{L}_{\text{sc}}^{\tilde{\mathbf{x}}\to\mathbf{x}}$. Therefore, the final *object-centric contrastive loss* function (ObjNCELoss) that we optimize is as follows:

$$\mathcal{L}_{\text{nce}}^{\mathbf{x}\leftrightarrow\tilde{\mathbf{x}}} = \mathcal{L}_{\text{nce}}^{\mathbf{x}\to\tilde{\mathbf{x}}} + \mathcal{L}_{\text{nce}}^{\tilde{\mathbf{x}}\to\mathbf{x}}. \tag{6}$$

3.4. Total Objective

To enhance the stability of the training process from the outset, we additionally employ a correspondence distillation loss [15], $\mathcal{L}_{\text{corr}}$ (see Supp D.1. for a detailed explanation). In total, we minimize the following objective $\mathcal{L}_{\text{total}}$:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{nce}}\mathcal{L}_{\text{nce}}^{\mathbf{x}\leftrightarrow\tilde{\mathbf{x}}} + (1-\lambda_{\text{nce}})\mathcal{L}_{\text{corr}} + \lambda_{\text{eig}}\mathcal{L}_{\text{eig}}, \tag{7}$$

where $0 \leqq \lambda_{\text{nce}} \leqq 1$ and $0 \leqq \lambda_{\text{eig}} \leqq 1$ are hyperparameters. Here, λ_{nce} starts from zero and increases rapidly, indicating the growing influence of $\mathcal{L}_{\text{nce}}^{\mathbf{x}\leftrightarrow\tilde{\mathbf{x}}}$ during training.

4. Experiments

In this section, we first discuss the implementation details, including dataset configuration, evaluation protocols, and detailed experimental settings. Then, we evaluate our proposed method, *EAGLE*, both qualitatively and quantitatively while making a fair comparison with existing state-of-the-art methods. We also demonstrate the effectiveness of our proposed method through an ablation study. See the supplementary material for additional details.

4.1. Experimental Settings

Implementation Details. We use DINO [6] pretrained vision transformer $\mathcal{F}$ which is kept frozen during the training process as in the prior works [15, 43]. The training sets are

Figure 6. Comparison between K-means and EiCue. The bottom row presents EiCue, highlighting its superior ability to capture subtle structural intricacies and understand deeper semantic relationships, which is not as effectively achieved by K-means.

resized and five-cropped to 244×244. For segmentation head $\mathcal{S}_\theta$, we use two layers of MLP with ReLU [15, 43], and for projection head $\mathcal{Z}_\xi$ we constructed a single linear layer [43]. All backbones employed an embedding dimension D_S and D_Z of 512. For the EiCue, we extract 4 eigenvectors from the eigenbasis $\mathbf{V}$. In the inference stage, we post-process the segmentation map with DenseCRF [15, 25, 43]. See supplement for more details.

Datasets. We evaluate on (1) COCO-Stuff [4], (2) Cityscapes [9], and (3) Potsdam-3 [20] datasets, in line with methodologies established in prior works [8, 15, 20, 43]. (1) The COCO-Stuff dataset is composed of its detailed pixel-level annotations, facilitating comprehensive various object understanding, while (2) Cityscapes presents diverse urban street scenes. (3) The Potsdam-3 dataset is composed of satellite imagery. Following the class selection protocols from previous studies [8, 15, 20, 43], we use 27 classes from both COCO-Stuff and Cityscapes. For Potsdam-3, we use all 3 classes (see supplement for result).

Evaluation Details. To align with established benchmarks, we adopt the evaluation protocols of prior works [15, 43]. Our evaluation includes (1) a linear probe, assessing representational quality with a supervised linear layer on the unsupervised model, and (2) clustering through semantic segmentation via minibatch K-means based on cosine distance [31], without ground truth, compared against it using Hungarian matching. We measure performance using pixel accuracy (Acc.) and mean Intersection over Union (mIoU).

4.2. Evaluation Results

Here, we carefully compare our proposed method to existing USS works in both qualitative and quantitative ways. We mainly set up two representative baselines [15, 43] from the literature which share the same evaluation protocols.

Table 1. Quantitative results on the COCO-Stuff dataset [4].

Method	Backbone	Unsupervised		Linear	
		Acc.	mIoU	Acc.	mIoU
DC [5]	R18+FPN	19.9	-	-	-
MDC [5]	R18+FPN	32.2	9.8	48.6	13.3
IIC [20]	R18+FPN	21.8	6.7	44.5	8.4
PiCIE [8]	R18+FPN	48.1	13.8	54.2	13.9
PiCIE+H [8]	R18+FPN	50.0	14.4	54.8	14.8
SlotCon [50]	R50	42.4	18.3	-	-
DINO [6]	ViT-S/16	22.0	8.0	50.3	18.1
+ STEGO [15]	ViT-S/16	52.5	23.7	70.6	34.5
+ HP [43]	ViT-S/16	54.5	24.3	74.1	39.1
+ *EAGLE (Ours)*	ViT-S/16	**60.1**	**24.4**	**75.2**	**42.5**
DINO [6]	ViT-S/8	28.7	11.3	68.6	33.9
+ TransFGU [52]	ViT-S/8	52.7	17.5	-	-
+ STEGO [15]	ViT-S/8	48.3	24.5	74.4	38.3
+ HP [43]	ViT-S/8	57.2	24.6	75.6	42.7
+ *EAGLE (Ours)*	ViT-S/8	**64.2**	**27.2**	**76.8**	**43.9**

Table 2. Quantitative results on the Cityscapes dataset [9].

Method	Backbone	Unsupervised		Linear	
		Acc.	mIoU	Acc.	mIoU
MDC [5]	R18+FPN	40.7	7.1	-	-
IIC [20]	R18+FPN	47.9	6.4	-	-
PiCIE [8]	R18+FPN	65.5	12.3	-	-
DINO [6]	ViT-S/8	34.5	10.9	84.6	22.8
+ TransFGU [52]	ViT-S/8	77.9	16.8	-	-
+ HP [43]	ViT-S/8	80.1	18.4	**91.2**	30.6
+ *EAGLE (Ours)*	ViT-S/8	**81.8**	**19.7**	**91.2**	**33.1**
DINO [6]	ViT-B/8	43.6	11.8	84.2	23.0
+ STEGO [15]	ViT-B/8	73.2	21.0	90.3	26.8
+ HP [43]	ViT-B/8	**79.5**	18.4	90.9	33.0
+ *EAGLE (Ours)*	ViT-B/8	79.4	**22.1**	**91.4**	**33.4**

Quantitative Evaluation: COCO-Stuff. In Table 1, our *EAGLE* method sets new benchmarks on the COCO-Stuff dataset. **(I)** With the ViT-S/8 backbone, *EAGLE* showcases substantial improvements over existing methods in unsupervised accuracy, with gains of **+15.9** over STEGO [15] and **+7.0** over HP [43]. The unsupervised mIoU of *EAGLE* also significantly outperforms other methods: **+2.7** over STEGO and **+2.6** over HP. The linear accuracy and mIoU of *EAGLE* both bring notable improvements over STEGO (**+2.4** Acc. and **+5.6** mIoU) and HP (**+1.2** Acc. and **+1.2** mIoU). Compared to SlotCon [50], which also emphasizes object-level representations, our model excels with a **+21.8** and **+8.9** in unsupervised mIoU and accuracy respectively. **(II)** With the ViT-S/16 backbone, *EAGLE* maintains its dominance, gaining **+7.6** over STEGO and **+5.6** over HP in unsupervised Acc. The linear accuracy and mIoU of *EAGLE* outperforms STEGO (**+4.6** Acc. and **+8.0** mIoU) and HP (**+1.1** Acc. and **+3.4** mIoU) as well.

Quantitative Evaluation: Cityscapes. As shown in Table 2, our evaluations on the Cityscapes dataset show that

Table 3. Ablation results on the COCO-Stuff dataset [4].

Exp. #	$\mathcal{L}_{\text{corr}}$	$\mathbf{x} \to \tilde{\mathbf{x}}$ $\mathcal{L}_{\text{obj}}$	$\mathbf{x} \to \tilde{\mathbf{x}}$ $\mathcal{L}_{\text{sc}}$	$\tilde{\mathbf{x}} \to \mathbf{x}$ $\mathcal{L}_{\text{obj}}$	$\tilde{\mathbf{x}} \to \mathbf{x}$ $\mathcal{L}_{\text{sc}}$	$\mathcal{M}_{\text{eicue}}$	$\mathcal{M}_{\text{km}}$	Unsupervised Acc.	Unsupervised mIoU
1	✓							46.9	21.8
2	✓		✓		✓	✓		59.3	23.2
3	✓	✓		✓		✓		62.1	25.1
4	✓			✓	✓	✓		61.6	24.8
5	✓	✓	✓			✓		62.9	26.1
6	✓	✓	✓	✓	✓		✓	55.1	17.0
7	✓	✓	✓	✓	✓	✓		**64.2**	**27.2**

EAGLE notably excels in both ViT-S/8 and ViT-B/8 backbones. **(I)** For the ViT-S/8 backbone, *EAGLE* has achieved significant unsupervised performance over STEGO (**+3.9** Acc. and **+2.9** mIoU) and HP (**+1.7** Acc. and **+1.3** mIoU). **(II)** For the ViT-B/8 backbone, *EAGLE* significantly improves both unsupervised Acc. and mIoU. The Cityscapes dataset innately exhibits highly imbalanced pixel-level class distributions, like the predominance of `sky` over `traffic light` pixels, typically forces a trade-off between Acc. and mIoU [43], as seen with STEGO and HP excelling in each metric respectively. However, *EAGLE* effectively balances these competing metrics, showcasing strong performance in both areas despite such challenges.

Qualitative Analysis. In Fig. 5, we also qualitatively compare our method to previous state-of-the-art models [15, 43] on the COCO-Stuff and Cityscapes datasets trained using ViT-S/8 and ViT-B/8 backbone, respectively. Our approach outperforms baselines by accurately segmenting objects and preserving details, unlike STEGO which tends to segment multiple elements within a single object `furniture` or `road`, and HP neglects certain small objects `sports(kite)` or `traffic sign`. Our model, however, is trained at the object level with an understanding of the structure of the image, which not only comprehends the overall layout but also ensures no objects are missed.

4.3. Ablation Study

We further analyze our model with ablation studies and discuss the results based on the full ablation results in Table 3 denoted as Exp. #1 to Exp. #7. We primarily conducted our experiments using the COCO-Stuff dataset using the DINO pretrained ViT-S/8 model. For more details, please refer to the supplementary material.

Effect of EiCue. We validate the effectiveness of EiCue ($\mathcal{M}_{\text{eicue}}$) by comparing the performance of our EiCue-enhanced method (Exp. #7) against a K-means ($\mathcal{M}_{\text{km}}$) approach (Exp. #6) in Table 3. The EiCue result shows a notable improvement, capturing fine structural details that K-means misses. Fig. 6 visually demonstrates how *EAGLE* better identifies object semantics and structures compared to K-means.

ObjNCE Loss. Table 3 shows how different loss components affect performance. The full model (Exp. #7) outperforms others, highlighting the effectiveness of combining all components. Notably, using $\mathcal{L}_{\text{obj}}$ alone (Exp. #3) significantly improves upon the baseline, underlining the importance of object-focused representation. The inclusion of $\mathcal{L}_{\text{sc}}$ further refines quality, as evidenced by comparing Exp. #3 with Exp. #7. Additionally, the combined use of both $\mathcal{L}_{\text{nce}}$ directions (Exp. #7) shows a synergistic effect over using them individually (Exp. #4 and Exp. #5).

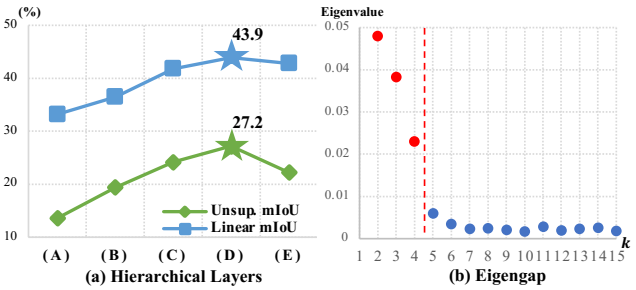

Figure 7. **(a)** Analysis of hierarchical attention with the following layer combinations (layer numbers in square brackets): (A): [1-6-12], (B): [12], (C): [11-12], (D): [10-11-12], and (E): [9-10-11-12]. **(b)** Analysis of eigengap to identify the optimal k for eigenbasis clustering, selected at the dashed line with maximal eigengap (i.e., the gap between two consecutive eigenvalues).

Combination of Hierarchical Attention and Eigengap. In Fig. 7a, we present results from using various combinations of hierarchical attention. The combination of the third-to-last, second-to-last, and last layers from 12-layer architecture, demonstrated the best performance since the layers closer to the end better capture the spatial information of the image. For optimal eigenbasis clustering, we conduct eigengap analysis in Fig. 7b. Since we choose k at the point where the eigengap is maximized, we have selected $k = 4$.

5. Conclusion

In this study, we present *EAGLE*, a novel method that addresses the persistent challenges in semantic segmentation with a focus on collecting semantic pairs through an object-centric lens. Through empirical analysis using a series of datasets, *EAGLE* showcases a remarkable capability to leverage the Laplacian matrix constructed from attention-projected features and fortified by an object-level prototype contrastive loss, which guarantees the accurate association of objects with their corresponding semantic pairs. Pioneering in utilizing dual advanced techniques, this method marks a substantial advance in addressing the constraints of patch-level representation learning found in previous research. Consequently, *EAGLE* emerges as a powerful framework for encapsulating the semantic and structural intricacies of images in contexts devoid of labels.

Acknowledgement. This work was supported in part by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government through the Ministry of Science and ICT (MSIT) under Grant RS-2023-00219019, and Institute of Information & Communications Technology Planning & Evaluation (IITP) Grant funded by MSIT (No. 2020-0-01361, Artificial Intelligence Graduate School Program (Yonsei University)).

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4981–4990, 2018. 1

[2] Inigo Alonso, Alberto Sabater, David Ferstl, Luis Montesano, and Ana C Murillo. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8219–8228, 2021. 1

[3] Mirko Paolo Barbato, Paolo Napoletano, Flavio Piccoli, and Raimondo Schettini. Unsupervised segmentation of hyperspectral remote sensing images with superpixels. *Remote Sensing Applications: Society and Environment*, 28:100823, 2022. 2

[4] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2018. 6, 7, 8

[5] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018. 7

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 2, 3, 6, 7

[7] Jeff Cheeger. A lower bound for the smallest eigenvalue of the laplacian. In *Problems in Analysis: A Symposium in Honor of Salomon Bochner (PMS-31)*, pages 195–200. Princeton University Press, 2015. 4

[8] Jang Hyun Cho, Utkarsh Mall, Kavita Bala, and Bharath Hariharan. Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16794–16804, 2021. 1, 2, 7

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 6, 7

[10] Zhijie Deng and Yucen Luo. Learning neural eigenfunctions for unsupervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 551–561, 2023. 3

[11] Nameirakpam Dhanachandra, Khumanthem Manglem, and Yambem Jina Chanu. Image segmentation using k-means clustering algorithm and subtractive clustering algorithm. *Procedia Computer Science*, 54:764–771, 2015. 2

[12] Lei Ding, Hao Tang, and Lorenzo Bruzzone. Lanet: Local attention embedding to improve the semantic segmentation of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 59(1):426–435, 2020. 1

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2

[14] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaeser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1341–1360, 2020. 1

[15] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T. Freeman. Unsupervised semantic segmentation by distilling feature correspondences. In *International Conference on Learning Representations*, 2022. 1, 2, 3, 4, 6, 7, 8

[16] Robert Harb and Patrick Knöbelreiter. Infoseg: Unsupervised semantic image segmentation with mutual information maximization. In *DAGM German Conference on Pattern Recognition*, pages 18–32. Springer, 2021. 1, 2

[17] Olivier J Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron Van den Oord, Oriol Vinyals, and Joao Carreira. Efficient visual pretraining with contrastive detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10086–10096, 2021. 3

[18] Olivier J Hénaff, Skanda Koppula, Evan Shelhamer, Daniel Zoran, Andrew Jaegle, Andrew Zisserman, João Carreira, and Relja Arandjelović. Object discovery and representation networks. In *European Conference on Computer Vision*, pages 123–143. Springer, 2022. 3

[19] Jyh-Jing Hwang, Stella X Yu, Jianbo Shi, Maxwell D Collins, Tien-Ju Yang, Xiao Zhang, and Liang-Chieh Chen. Segsort: Segmentation by discriminative sorting of segments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7334–7344, 2019. 2

[20] Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9865–9874, 2019. 1, 2, 7

[21] Feng Jiang, Aleksei Grigorev, Seungmin Rho, Zhihong Tian, YunSheng Fu, Worku Jifara, Khan Adil, and Shaohui Liu. Medical image semantic segmentation based on deep learning. *Neural Computing and Applications*, 29:1257–1265, 2018. 1

[22] Robin Karlsson, Tomoki Hayashi, Keisuke Fujii, Alexander Carballo, Kento Ohtani, and Kazuya Takeda. Improving dense representation learning by superpixelization and contrasting cluster assignment. In *British Machine Vision Conference*, 2021. 1, 2

[23] Hoel Kervadec, Jose Dolz, Meng Tang, Eric Granger, Yuri Boykov, and Ismail Ben Ayed. Constrained-cnn losses for weakly supervised segmentation. *Medical Image Analysis*, 54:88–99, 2019. 1

[24] Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Mean shift for self-supervised learning.

In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10326–10335, 2021. 2

[25] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in Neural Information Processing Systems*, 24, 2011. 7

[26] Hyeokjun Kweon, Sung-Hoon Yoon, and Kuk-Jin Yoon. Weakly supervised semantic segmentation via adversarial learning of classifier and reconstructor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11329–11339, 2023. 1

[27] Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Semi-supervised semantic segmentation with directional context-aware consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1205–1214, 2021. 1

[28] Rui Li, Shunyi Zheng, Ce Zhang, Chenxi Duan, Jianlin Su, Libo Wang, and Peter M Atkinson. Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2021. 1

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2

[30] Ming Luo, Yu-Fei Ma, and Hong-Jiang Zhang. A spatial constrained k-means approach to image segmentation. In *Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint*, pages 738–742. IEEE, 2003. 2

[31] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297. Oakland, CA, USA, 1967. 5, 7

[32] R Manavalan and K Thangavel. Trus image segmentation using morphological operators and dbscan clustering. In *2011 World Congress on Information and Communication Technologies*, pages 898–903. IEEE, 2011. 2

[33] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8364–8375, 2022. 3, 5

[34] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, 14, 2001. 2, 4

[35] Pedro O O Pinheiro, Amjad Almahairi, Ryan Benmalek, Florian Golemo, and Aaron C Courville. Unsupervised learning of dense visual representations. *Advances in Neural Information Processing Systems*, 33:4489–4500, 2020. 3

[36] Yassine Ouali, Céline Hudelot, and Myriam Tami. Autoregressive unsupervised image segmentation. In *European Conference on Computer Vision*, pages 142–158. Springer, 2020. 1, 2

[37] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020. 1

[38] Thrasyvoulos N Pappas and Nikil S Jayant. An adaptive clustering algorithm for image segmentation. In *International Conference on Acoustics, Speech, and Signal Processing,*, pages 1667–1670. IEEE, 1989. 2

[39] Shenghai Rong, Bohai Tu, Zilei Wang, and Junjie Li. Boundary-enhanced co-training for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19574–19584, 2023. 1

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 1

[41] Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, and Francesco Locatello. Bridging the gap to real-world object-centric learning. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[42] Ramprasaath R Selvaraju, Karan Desai, Justin Johnson, and Nikhil Naik. Casting your model: Learning to localize improves self-supervised representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11058–11067, 2021. 3

[43] Hyun Seok Seong, WonJun Moon, SuBeen Lee, and Jae-Pil Heo. Leveraging hidden positives for unsupervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19540–19549, 2023. 1, 2, 3, 4, 6, 7, 8

[44] Jianbing Shen, Xiaopeng Hao, Zhiyuan Liang, Yu Liu, Wenguan Wang, and Ling Shao. Real-time superpixel segmentation by dbscan clustering algorithm. *IEEE Transactions on Image Processing*, 25(12):5933–5942, 2016. 2

[45] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000. 2, 4

[46] Marvin Teichmann, Michael Weber, Marius Zoellner, Roberto Cipolla, and Raquel Urtasun. Multinet: Real-time joint semantic reasoning for autonomous driving. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1013–1020. IEEE, 2018. 1

[47] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10052–10062, 2021. 3

[48] Alexander Vezhnevets and Joachim M Buhmann. Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3249–3256. IEEE, 2010. 1

[49] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021. 3

[50] Xin Wen, Bingchen Zhao, Anlin Zheng, Xiangyu Zhang, and Xiaojuan Qi. Self-supervised visual representation learning with semantic grouping. In *Advances in Neural Information Processing Systems*, 2022. 3, 7

[51] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021. 3

[52] Zhaoyuan Yin, Pichao Wang, Fan Wang, Xianzhe Xu, Hanling Zhang, Hao Li, and Rong Jin. Transfgu: a top-down approach to fine-grained unsupervised semantic segmentation. In *European Conference on Computer Vision*, pages 73–89. Springer, 2022. 1, 2, 7

[53] Andrii Zadaianchuk, Matthaeus Kleindessner, Yi Zhu, Francesco Locatello, and Thomas Brox. Unsupervised semantic segmentation with self-supervised object-centric representations. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[54] Xiao Zhang and Michael Maire. Self-supervised visual representation learning from hierarchical grouping. *Advances in Neural Information Processing Systems*, 33:16579–16590, 2020. 3

[55] Xiaosong Zhang, Yunjie Tian, Lingxi Xie, Wei Huang, Qi Dai, Qixiang Ye, and Qi Tian. Hivit: A simpler and more efficient design of hierarchical vision transformer. In *The Eleventh International Conference on Learning Representations*, 2023. 2