

Deep Spectral Methods: A Surprisingly Strong Baseline for Unsupervised Semantic Segmentation and Localization

Luke Melas-Kyriazi

Christian Rupprecht

Iro Laina

Andrea Vedaldi

University of Oxford

{lukemk, chrissr, iro, vedaldi}@robots.ox.ac.uk

Abstract

*Unsupervised localization and segmentation are long-standing computer vision challenges that involve decomposing an image into semantically meaningful segments without any labeled data. These tasks are particularly interesting in an unsupervised setting due to the difficulty and cost of obtaining dense image annotations, but existing unsupervised approaches struggle with complex scenes containing multiple objects. Differently from existing methods, which are purely based on deep learning, we take inspiration from traditional spectral segmentation methods by reframing image decomposition as a graph partitioning problem. Specifically, we examine the eigenvectors of the Laplacian of a feature affinity matrix from self-supervised networks. We find that these eigenvectors already decompose an image into meaningful segments, and can be readily used to localize objects in a scene. Furthermore, by clustering the features associated with these segments across a dataset, we can obtain well-delineated, nameable regions, i.e. semantic segmentations. Experiments on complex datasets (PASCAL VOC, MS-COCO) demonstrate that our simple spectral method outperforms the state-of-the-art in unsupervised localization and segmentation by a significant margin. Furthermore, our method can be readily used for a variety of complex image editing tasks, such as background removal and compositing.*¹

1. Introduction

Well-established computer vision tasks such as localization and segmentation are aimed at understanding the structure of images at a fine level of detail. The modern approach to these tasks, which consists of training deep neural networks in an end-to-end fashion, has shown strong performance given large quantities of human-labeled data. How-

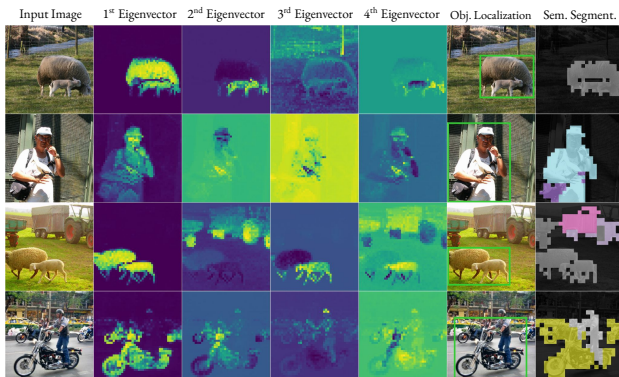


Figure 1. **Deep Spectral Methods.** We present a simple approach based on spectral methods that decomposes an image using the eigenvectors of a Laplacian matrix constructed from a combination of color information and unsupervised deep features. The method surpasses the state of the art in unsupervised image segmentation and object localization while also being significantly simpler.

ever, obtaining labeled data for these dense tasks can be difficult and expensive. Whereas vast quantities of (weak) image labels and descriptions may be obtained from the Internet [43, 55, 63], dense image annotations cannot be easily sourced from it, and creating them manually is a labor-intensive process. Moreover, in many specialized fields such as medical imaging, where detection and segmentation tasks are particularly important, data labeling must be manually performed by a domain expert. As a result, performing dense vision tasks without labeled data is an important open problem.

Numerous existing methods perform dense visual tasks with weak annotations, such as image-level labels, captions, spoken narratives, scribbles, and points [9, 36, 46, 47]. However, fully-unsupervised dense image understanding remains challenging and under-explored. Current approaches involve clustering dense features (e.g., IIC [41]), contrastive learning with saliency masks (e.g., MaskContrast [74]), and

¹Project Page: <https://lukemelas.github.io/deep-spectral-segmentation/>

GAN-based object discovery (*e.g.*, ReDo [13]). Others segment images of a given object category into a number of semantic regions (parts) [20, 22, 37]. However, these approaches tend to struggle with complex images, and most of them can only identify a single object per image.

In this paper, we take inspiration from image segmentation methods from the pre-deep learning era, which framed the segmentation problem as one of graph partitioning. Whereas existing unsupervised segmentation methods are based primarily on deep learning, we show the benefits of combining deep learning with traditional graph-theoretic methods.

Our method first utilizes a self-supervised network to extract dense features corresponding to image patches. We then construct a weighted graph over patches, where edge weights give the semantic affinity of pairs of patches, and we consider the eigendecomposition of this graph’s Laplacian matrix. We find that without imposing any additional structure, the eigenvectors of the Laplacian of this graph directly correspond to semantically meaningful image regions. Notably, the eigenvector with the smallest nonzero eigenvalue generally corresponds to the most prominent object in the scene. We show that, surprisingly, simply extracting bounding boxes or masks from this eigenvector surpasses the current state of the art on unsupervised object localization/segmentation across numerous benchmarks.

Next, we propose a pipeline for semantic segmentation. We first convert the eigensegments into discrete image regions by thresholding and associate each region with a semantic feature vector from the network. We consider all image regions in a large dataset of images and jointly cluster these regions, yielding semantic (pseudo-)labels that are consistent across the dataset. Lastly, we perform self-training using these labels to refine our results, and we evaluate against the ground truth segmentations. Different from prior self-supervised semantic segmentation methods, our method performs well on complex images without fine-tuning. Importantly, while recent GAN-based and saliency-based methods are limited to finding a single semantic region per image, our method can segment multiple semantic regions and outperforms prior methods on PASCAL VOC 2012, re-surfacing spectral methods as a strong baseline for future work.

Finally, we show that a slight variant of our method is well-suited to the task of soft image decomposition (*i.e.*, image matting), or breaking down an (RGB) image into multiple (RGB-A) layers with soft boundaries. This decomposition dramatically simplifies real-world image editing and compositing tasks such as background replacement.

2. Related work

In this section, we discuss the different fields related to our approach along with the relatively new tasks of unsu-

pervised localization and segmentation.

Self-Supervised Visual Representation Learning The past five years have seen tremendous progress in self-supervised visual representation learning. Early research in this area was based on solving pretext tasks such as rotation, jigsaw puzzles, colorization, and inpainting [24, 27, 39, 59, 61, 62, 95, 96]. Recent methods mostly consist of contrastive learning with heavy data augmentation [15, 16, 30, 32, 34, 44, 61, 69, 70, 87]. Others distinguish themselves by removing the dependency on negative examples [17, 29, 91], the use of clustering [7, 52], and most recently, the extension to transformer architectures [8, 18].

As these methods primarily focused on the downstream task of image classification, they were designed to produce a single global feature vector for each input image. Although numerous specialized methods have been proposed for dense contrastive learning [60, 83, 88], they require fine-tuning for application on tasks such as detection or segmentation. Recently, however, the emergence of self-supervised vision transformers has made it possible to extract dense (patch-wise) feature vectors without the need for specialized dense contrastive learning methods.

Vision Transformers The past year in computer vision research has been marked by the emergence of the vision transformer (ViT) architecture, a variant of the transformer model popular in sequence modeling and NLP [75]. Vision transformers reshape an image into a sequence of small patches and apply layers of self-attention before aggregating the result into an additional global token, usually denoted [CLS]. Numerous variants of the vision transformer have been proposed, such as DeiT [71], Token-to-Token ViT [90], DeepViT [98], CrossViT [11], PiT [35], CaiT [72], LeViT [28], CvT [85], and Swin/Twins [21, 54]. Nonetheless, the defining aspect of this class of models is the use of self-attention to process information throughout the network. Since self-attention involves self-comparisons of image patch features, it is natural to construct a semantic affinity matrix over patches, as we do in our approach.

Numerous recent works [4, 8, 18] have observed that vision transformers perform exceptionally well relative to convolutional architectures in the self-supervised setting. In particular, Caron *et al.* [8] train a self-supervised ViT using multi-crop training and a momentum encoder, and argue that it learns better-localized features than supervised ViTs or ResNets. They show that the self-attention layer of the last [CLS] token has heads that localize foreground objects in the scene. This is a baseline in our object localization experiments, upon which we improve substantially.

Unsupervised Localization Unsupervised localization refers to the task of finding the location of an object in the form of a bounding box. Most approaches to localization find objects by identifying co-occurring patterns across image collections [19, 76, 77]. However, due to inter-image

comparisons, these approaches are generally slow, memory-intensive, and have trouble scaling to large datasets.

Recently, some works have begun to eschew this paradigm by finding objects purely from the features of pre-trained deep networks. [22] perform co-localization through deep feature factorization. [93] mines patterns from pre-trained convolutional network features. Most recently, LOST [67] extracts localization information from the attention features of a self-supervised vision transformer. LOST uses the same features as we do in our approach, but extracts bounding boxes using seed selection and expansion based on the number of patch correlations. By contrast, our approach localizes objects through spectral bipartitioning, which is more principled and more performant than [67].

Unsupervised Segmentation Whereas a plethora of methods have tackled segmentation in semi- and weakly-supervised settings [9, 36, 46, 47], the unsupervised setting remains a relatively nascent area of research. Current methods can broadly be characterized as either generative or discriminative. Generative methods use unlabeled images to train special-purpose image generators [1, 3, 13], or directly extract segmentations from pretrained generators [56, 80]; discriminative methods are primarily based on clustering and contrastive learning [38, 41, 74, 97]. MaskContrast [74], the current state of the art in unsupervised semantic segmentation, uses saliency detection to find object segments (*i.e.*, the foreground) and then learns pixel-wise embeddings via a contrastive objective. However, MaskContrast relies heavily on a saliency network which is initialized with a pretrained (fully-supervised) network. Moreover, it heavily relies on the assumption that all foreground pixels in a given image belong to the same object category, which fails to hold for most images containing multiple objects.

Spectral Methods Spectral graph theory emerged from the study of continuous operators on Riemannian manifolds [10]. Subsequent works brought this line of research to the discrete setting of graphs, with numerous results connecting global properties of graphs to the eigenvalues and eigenvectors of their Laplacian matrices. Two of these results are essential for understanding our work. First, [26] showed that the second-smallest eigenvalue of a graph, now called the algebraic connectivity or the Fiedler eigenvalue, quantifies the connectivity of a graph. In our work, we use the Fiedler eigenvector for object localization. Second, [25, 26] showed that the eigenvectors of graph Laplacians yield minimum-energy graph partitions. In our work, we use this idea to extract semantic segmentations from a patch-wise *semantic* affinity matrix.

Spectral graph methods were popularized within the machine learning and computer vision communities by [57, 64]. Shi and Malik [64] framed image segmentation in the language of graph cuts, that is finding a partition of the graph (*i.e.* the image) to minimize the similarity of the

partitions. They noted that finding the minimum cut generally leads to smaller-than-desired partitions, so they normalized by the total weight of all edges connected to each partition. Ng *et al.* [57] also performed spectral decomposition, and then stacked and clustered the eigenvectors along the eigenvector dimension in order to obtain a fixed number of partitions. Numerous follow-up works have extended these ideas to new settings such as co-segmentation [86].

3. Method

Our method is rooted in ideas from spectral graph theory, the study of analyzing properties of graphs by looking at their spectra (*i.e.* their eigenvalues). We begin with a general overview of spectral methods.

3.1. Background

Let $G = (V, E)$ be a weighted undirected graph with adjacency matrix $W = \{w(u, v) : (u, v) \in E\}$. The Laplacian matrix L of this graph is given by $L = D - W$ or in the normalized case $L = D^{-1/2}(D - W)D^{-1/2}$, where D is the diagonal matrix whose entries contain the row-wise sum of W . The Laplacian is particularly interesting because it corresponds to a quadratic form

$$x^T L x = \sum_{(u,v) \in E} w(u, v) \cdot (x(u) - x(v))^2$$

for $x \in \mathbb{R}^n$ with $n = |V|$. Intuitively, this quadratic form measures the smoothness of a function x defined on the graph G . If x is smooth with respect to G , then $x(u)$ is similar to $x(v)$ whenever u is similar to v (where similarity is quantified by the weight $w(u, v)$).

The eigenvectors and eigenvalues of L are the central objects of study in spectral graph theory. The eigenvectors y_i span an orthogonal basis for functions on G that is, in the sense described above, the *smoothest* possible orthogonal basis:

$$y_i = \operatorname{argmin}_{\|x\|=1, x \perp y_{<i}} x^T L x$$

with $y_0 = \mathbf{1} \in \mathbb{R}^n$. The eigenvalues λ_i , with $0 = \lambda_0 \leq \lambda_1 \leq \dots \leq \lambda_{n-1}$, are the values of the right-hand side of the above expression (which is known as the Courant-Fischer theorem in the spectral graph theory community). Thus, it is natural to express functions on G in the basis of the eigenvectors of the graph Laplacian.

In the context of classical image segmentation, the nodes of G correspond to the pixels of an image $I \in \mathbb{R}^{MN}$, and the edge weights $W \in \mathbb{R}^{MN \times MN}$ correspond to the affinities between pairs of pixels. The eigenvectors $y \in \mathbb{R}^{MN}$ can be thought of as soft image segments. Discrete graph partitions are often called graph cuts, where the value of cut with two partitions A and B ($A \cup B = V$, $A \cap B = \emptyset$) is given by

$$\operatorname{cut}(A, B) = \sum_{u \in A, v \in B} w(u, v)$$

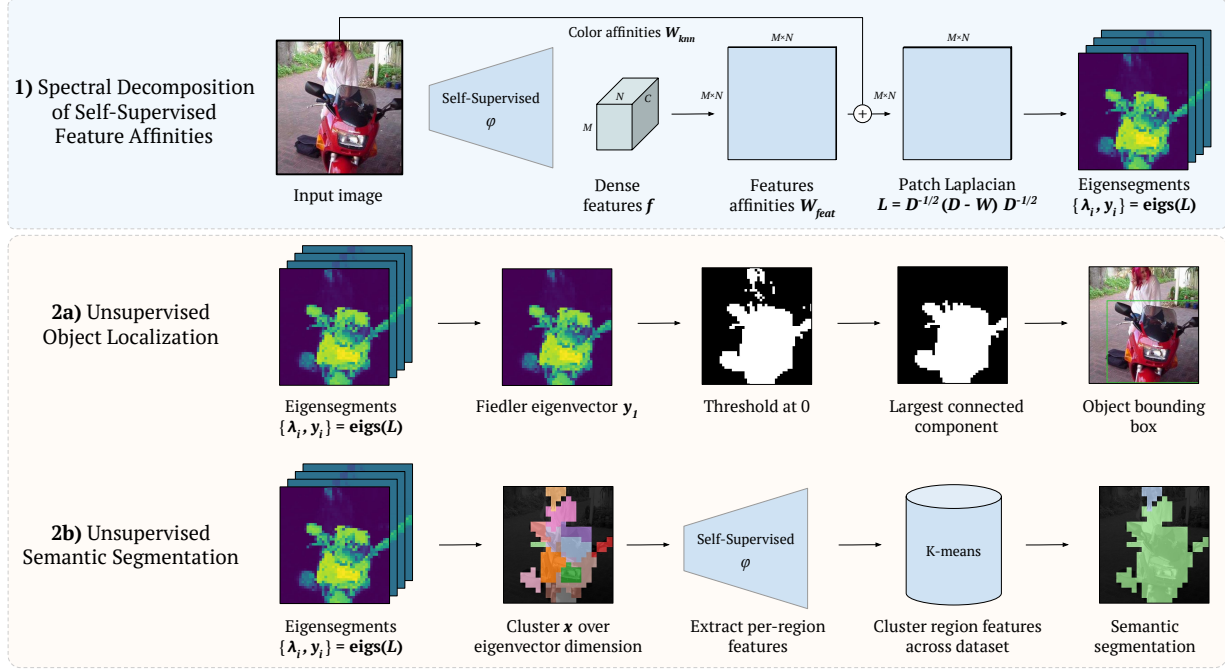


Figure 2. **Method overview.** Our approach to unsupervised segmentation combines the benefits of modern self-supervised learning with the benefits of traditional spectral methods. Given an image, we first extract dense features from a network ϕ and use these to construct a semantic affinity matrix, which we then fuse with low-level color information. We decompose the image into soft segments by computing the eigenvectors of the Laplacian of this matrix. Second, we can use these eigenvectors for a wide range of downstream tasks. For object localization (2a), we find that simply taking the sign of the eigenvector with the smallest nonzero eigenvalue, and placing a bounding box around this region, produces state-of-the-art object localization performance. For semantic segmentation (2b), we convert the eigenvectors into discrete segments, compute a feature vector for each segment, and cluster these segments across an entire dataset.

i.e., the total weight of edges that have been removed by partitioning. A normalized version of graph cuts [65] is particularly useful in practice, and emerges naturally from the eigenvectors of the normalized Laplacian; the optimal (continuous) bi-partitioning in this case reduces to finding the Laplacian eigenvectors.

The choice of W The most important aspect of spectral methods for image segmentation is the construction of the adjacency matrix W , which encodes the similarities of each pair of pixels.

To take the spatial neighborhood and color information into account, some works (especially for image matting [14, 51]) use a nearest neighbor formulation. KNN-matting converts an image to the HSV color space, and defines for each pixel a vector $\psi(u) = (\cos(c_H), \sin(c_H), c_S, c_V, p_x, p_y) \in \mathbb{R}^5$ containing both color information (the c_H, c_S, c_V values) and spatial information (the p_x, p_y values). They then construct a sparse affinity matrix from pixel-wise nearest neighbors based on ψ :

$$W_{\text{knn}}(u, v) = \begin{cases} 1 - \|\psi(u) - \psi(v)\|, & u \in \text{KNN}_{\psi}(v), \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where $u \in \text{KNN}_{\psi}(v)$ are the k -nearest neighbors of v under the distance defined by ψ .

This affinity matrix typically yields sharp object boundaries, but since it is designed for the image matting setting, in which a trimap is given, it does not take semantic information into account. While some recent works have explored the idea of semantic matting [2, 68], they tackle a different problem from us and are limited by their use of large labeled datasets. Our work combines the benefits of this classical spectral decomposition with the benefits of deep self-supervised features for the purpose of unsupervised object and scene understanding.

3.2. Semantic Spectral Decomposition

In this section we will describe our formulation of deep spectral decomposition and its direct application to the down-stream tasks of object localization and semantic segmentation. An overview is shown in Fig. 2.

Let $I \in \mathbb{R}^{3 \times M \times N}$ be an image. We will first decompose I into semantically consistent regions using features of a neural network. These regions can then be processed into bounding boxes for object localization or clustered across a collection of images for semantic segmentation.

We begin by extracting deep features $f = \phi(I) \in \mathbb{R}^{C \times M/P \times N/P}$ using a network ϕ . As the feature maps of a network are usually computed at a lower resolution than the image, we introduce P for this downsampling factor. In the case where ϕ is a transformer architecture, P represents the patch size. These features may be extracted from any part of the network and can even be a combination of different layers, *i.e.*, hyper-columns [31]. In our experiments with transformers, we find, similarly to [67], that features from the keys of the last attention layer work especially well, as they are inherently meant for self-aggregation of similar features.

We then construct an affinity matrix from the patchwise feature correlations. Additionally, we threshold the affinities at 0, because the features are designed for aggregating similar features rather than anti-correlated features:

$$W_{\text{feat}} = f f^T \odot (f f^T > 0) \in \mathbb{R}^{\frac{MN}{P^2} \times \frac{MN}{P^2}} \quad (2)$$

These feature affinities contain rich semantic information at a coarse resolution. To gain back low-level details, we fuse them with traditional color-level information which can be seen as features from the 0-th layer of the network.

To perform the fusion, we bilinearly upsample the features and downsample the image to an intermediate resolution $M' \times N'$. Empirically, we find that using $M' = M/8$ and $N' = N/8$ yields good low-level details while maintaining a fast runtime. As our color affinity matrix, we use the sparse KNN-matting matrix (Eq. (1)), although any traditional similarity matrix can also be used. Our final affinity matrix is the weighed sum of the feature and color matrices:

$$W = W_{\text{feat}} + \lambda_{\text{knn}} W_{\text{knn}} \quad (3)$$

where λ_{knn} is a user-defined parameter that trades off semantic and color consistency.

Given W , we take the eigenvectors of its Laplacian $L = D^{-1/2}(D - W)D^{-1/2}$ to decompose an image into soft segments: $\{y_0, \dots, y_{n-1}\} = \text{eigs}(L)$. Since the first eigenvector y_0 , is a constant vector corresponding to $\lambda_0 = 0$, for our purposes we use only $y_{>0}$.

In Fig. 3, we show qualitative examples of our eigendecomposition. We find that the eigensegments correspond to semantically meaningful image regions and have well-delineated boundaries. As such, localization and segmentation tasks arise as natural immediate applications of this approach. Importantly, if the choice of ϕ is a model trained with self-supervision on unlabeled data, then it is possible to address these tasks without supervision and without the need for finetuning.

3.3. Object Localization

Object localization is the task of identifying, in the form of a bounding box, the location of the primary object in

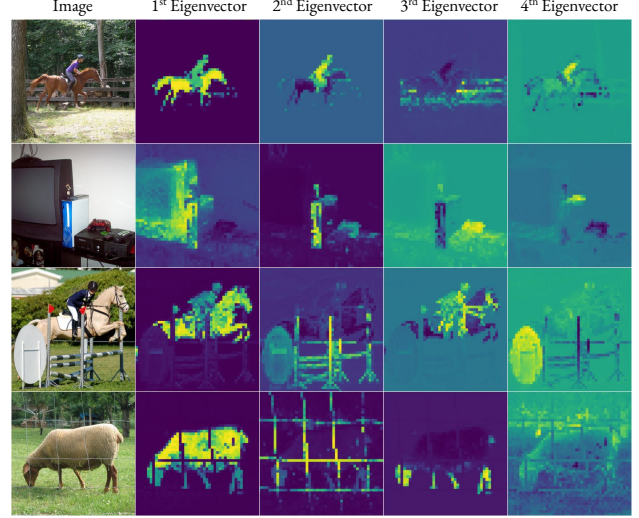


Figure 3. **Eigenvectors on PASCAL VOC 2012.** Examples of images and the first 4 corresponding eigenvectors of our feature affinity matrix (excluding the zero-th constant eigenvector). The eigenvectors correspond to semantic regions, with the first eigenvector usually identifying the most salient object in the image.

an image. To localize the main object, we simply follow the standard spectral bisection approach. We examine the Fiedler eigenvector y_1 of L and discretize it by taking its sign to obtain a binary segmentation of an image. We then take a bounding box around the smaller of the two regions, which is more likely to correspond to any foreground object rather than the background.

3.4. Object Segmentation

Object segmentation, also called foreground-background segmentation, is a binary segmentation task that is very closely related to object localization; it consists of densely segmenting the foreground object in an image. We approach this task in the same manner as object localization, by using the Fiedler eigenvector y_1 to first find a coarse object segmentation. However, rather than taking a bounding box around our coarse segment, we apply a simple pairwise CRF to increase the resolution of our segmentation from $M' \times N'$ to $M \times N$.

3.5. Semantic Segmentation

Semantic segmentation is the task of categorizing each pixel in an image with a label that is semantically consistent across an entire dataset. Ideally, we would approach this problem by constructing a Laplacian matrix L of size $MNT \times MNT$ containing the pairwise affinities of all pixels in an entire dataset of size T , but this is computationally infeasible. As a result, we perform a three-step process in which we: (1) break each image into segments, (2) compute a feature vector for each segment, and (3) cluster these

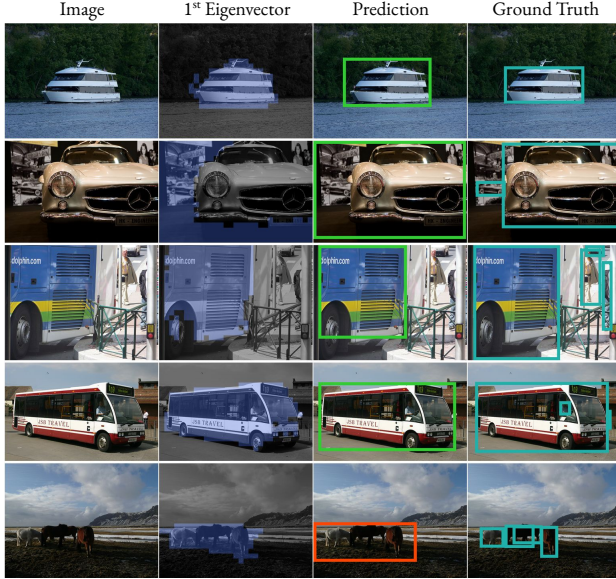


Figure 4. **Object localization on PASCAL VOC 2012.** From left to right, we show the original image, a visualization of the mask produced by thresholding the Fiedler eigenvector at 0, our predicted bounding boxes, and the ground truth bounding boxes. We color our bounding boxes in green or red based on whether they do or do not have IoU greater than 50% with one of the ground-truth boxes.

segments across a collection of images.

For step (1), we discretize the first m eigenvectors $\{y_1, \dots, y_m\}$ of L by clustering them across the eigenvector dimension using K -means clustering (for every image separately). Since we do not know a-priori the number of meaningful segments for an image, we find empirically that it is preferable to over-cluster the image into regions. For step (2), we take a crop around each segment, and compute its feature vector f_s using our self-supervised transformer. For step (3), we cluster the set of all feature vectors $\{f_s\}$ across all images using K -means clustering with k clusters. The second clustering step assigns a label to each segment of each image; adjacent regions with the same label are merged together, effectively reducing the number of segments found per image as needed. At the end of this process, we arrive at a set of semantic image segmentations consistent across the entire dataset.

Finally, we apply a self-training step to further improve segmentation quality. We train a standard segmentation model with a self-supervised backbone using the segmentations obtained above as pseudolabels. As shown in Fig. 5, this distillation process improves segmentation quality and inter-image consistency.

Method	VOC-07	VOC-12	COCO-20k
Selective Search [78]	18.8	20.9	16.0
EdgeBoxes [73]	31.1	31.6	28.8
Kim et al. [48]	43.9	46.4	35.1
Zhang et al. [94]	46.2	50.5	34.8
DDT+ [84]	50.2	53.1	38.2
rOSD [99]	54.5	55.3	48.5
LOD [79]	53.6	55.1	48.5
DINO- [CLS] [8]	45.8	46.2	42.1
LOST [67]	61.9	64.0	50.7
Ours	62.7	66.4	52.2

Table 1. **Single-object localization performance (CorLoc).** As is standard practice, we use the trainval sets of PASCAL VOC 2007 and 2012 for evaluation.

4. Experiments

To evaluate the effectiveness of spectral methods for unsupervised tasks, we perform experiments on unsupervised object localization, object segmentation, semantic segmentation and image matting. We use DINO-ViT-Base [8] as our self-supervised transformer for feature extraction unless otherwise noted. Further implementation details are discussed in the supplementary material.

4.1. Object Localization

We compare our unsupervised object localization method to prior work on three benchmarks: PASCAL VOC 2007, PASCAL VOC 2012, and COCO-20k (a subset of 20k images from the MS-COCO dataset [53], introduced in [77]). We evaluate following the same procedure as [67, 78]. As is standard practice, we evaluate on the *trainval* sets of these datasets. Results are measured in terms of the Correct Localization (CorLoc) metric, which measures the percentage of images on which an object is correctly localized. Since the images typically contain multiple objects, a prediction bounding box is considered to have correctly identified an object if it has greater than 50% intersection-over-union with *any* ground truth bounding box.

We give quantitative results in Table 1. Compared to older methods based on identifying co-occurrences across image collections, our method delivers dramatically improved performance. Relative to the state-of-the-art, LOST [67], which uses the same self-supervised features, our method still meaningfully outperforms across all datasets. In Tab. 2 we ablate the influence of the architecture on the localization performance and find that transformer-based models score significantly higher than ResNets, with larger models performing better. Fig. 4 shows qualitative examples of our method.

In the supplementary material, we conduct a series of ablation studies to better understand our proposed method.

Model	Pretraining	LOST [67]	Ours
ResNet-50	DINO	36.8	26.9
ViT-S-16	MoCo-v3	32.5	37.3
ViT-S-16	DINO	61.9	61.6
ViT-B-16	MoCo-v3	53.3	61.1
ViT-B-16	DINO	60.1	61.6
ViT-S-8	DINO	55.5	62.6
ViT-B-8	DINO	46.6	62.7

Table 2. **Architecture and pretraining.** Single-object localization performance on PASCAL VOC 2007. All models are trained with $\lambda_{\text{km}} = 0$ to isolate the quality of the features. Vision transformers outperform the convolutional ResNet-50 model, possibly due to the inherent structure of self-attention layers.

Method	CUB	DUTS	OMR	ECSSD
[6] PertGAN	0.380	-	-	-
[13] ReDO	0.426	-	-	-
[45] UIBS	0.442	-	-	-
[42] IIC-seg	0.365	-	-	-
[5] OneGAN	0.555	-	-	-
[80] Voynov <i>et al.</i>	0.683	0.498	0.453	0.672
[56] Melas-Kyriazi <i>et al.</i>	0.664	0.528	0.509	0.713
<i>Ours</i>	0.769	0.514	0.567	0.733

Table 3. **Single-Object Segmentation.** We present mIoU scores across four datasets, comparing to highly-tuned methods based on GANs [5, 6, 13, 56, 80] and contrastive learning [42].

Our ablations generally support the idea that the spatial attention mechanism in vision transformers results in the learning of well-localized intermediate features. Concretely, we find that the most effective features for our tasks are those extracted from the attention keys in the later blocks of vision transformers.

4.2. Object Segmentation

For single-object segmentation, we evaluate on four challenging benchmarks: CUB [81], DUTS [82], DUT-OMRON [89], and ECSSD [66]. We evaluate in the same manner as [56, 80], using the standard mean intersection-over-union (mIoU) metric to measure performance. Quantitative results are shown in Tab. 3. Our approach generally outperforms highly-tuned methods based on layer-wise GAN learning (*i.e.*, PertGAN [6], ReDO [13], OneGAN [5]), extracting segmentations from GANs (*i.e.* [56, 80]), and dense contrastive training (*e.g.* IIC [42]). Qualitative results are shown in the supplement.

Method	mIoU
<i>Pretext task methods</i>	
Co-Occurrence [40]	4.0
CMP [92]	4.3
Colorization [95]	4.9
<i>Clustering/Contrastive methods</i>	
IIC [41]	9.8
MaskContrast [†] [74]	35.0
<i>Additional baselines</i>	
Cluster-Patch	5.3
Cluster-Seg	12.1
Saliency-DINO-ViT-B [†]	30.1
MaskContrast-DINO-ViT-B [†]	31.2
<i>Ours w/o self-training</i>	30.8 \pm 2.7
<i>Ours</i>	37.2 \pm 3.8

Table 4. **Semantic Segmentation on PASCAL VOC 2012.** We calculate the average and standard deviation of our method over four random seeds. [†] indicates methods using saliency networks that were initialized from supervised models.

4.3. Semantic Segmentation

We now consider the challenging setting of unsupervised semantic segmentation, which differs from object segmentation in that it involves identifying multiple semantic masks per image and associating these masks across images. We evaluate our approach on the PASCAL VOC 2012 dataset, which contains 21 semantic classes (20 classes and a background class). For the segmentation phase, we start by clustering the eigenvectors into 15 segments per image. We then compute a feature vector for each of these segments, and cluster these across the dataset using K -means with $k = 21$ clusters. For the self-training stage, we train a DeepLab [12] model with a ResNet-50 [33] backbone that is pre-trained with self-supervision using DINO. We freeze the first two layers of the ResNet backbone and train the rest using the AdamW [49] optimizer with learning rate of 0.005 for 3000 steps. We perform self-training on the training set with a standard cross-entropy loss. Finally, we evaluate the outputs of the final self-trained model on the validation set using mIoU. Since clustering results in pseudo-labeling of image segments, Hungarian matching [50] is used to optimally match clusters to ground truth labels.

In Tab. 4, we compare to prior methods, including the state-of-the-art MaskContrast [74]. We note that MaskContrast uses saliency masks from Deep-USPS [58], which was pre-trained with supervision using labeled data (Cityscapes [23]). For a fairer comparison, we also train a version of MaskContrast based on a DINO-pretrained model (MaskContrast-DINO-ViT-B). Additionally, we give results for directly clustering DINO-pretrained features masked with Deep-USPS saliency maps (Saliency-DINO-

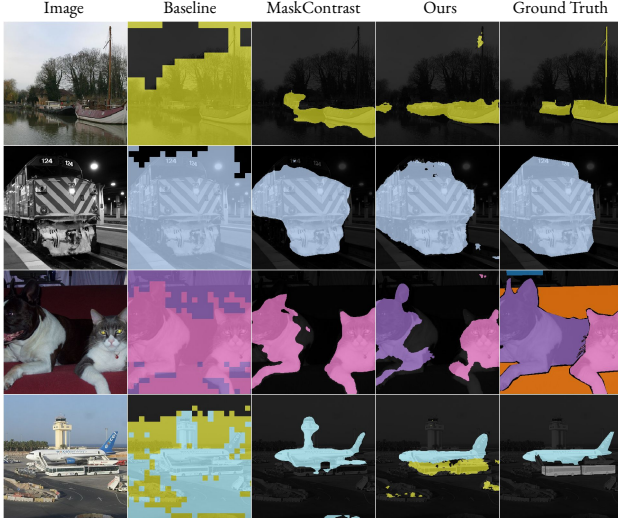


Figure 5. **Sematic Segmentation on PASCAL VOC 2012.** From left to right: we show the original image, the Cluster-Seg baseline, MaskContrast [74] segmentation, our results and the ground truth. The Cluster-Seg baseline often fails to adequately capture object boundaries. Whereas MaskContrast is only able to identify a single semantic class per image, our method successfully segments multiple semantic objects in the same image.

ViT-B), *i.e.*, without the additional dense training; this alone almost reaches the performance of MaskContrast.

Finally, we compare to two additional clustering-based baselines. The first baseline (Cluster-Patch) is to directly cluster the self-supervised image features for all patches across the entire dataset. That is, we extract $\frac{MN}{P^2} \cdot T$ features from all T images and assign them to clusters via K -means with 21 clusters. The second baseline (Cluster-Seg) uses K -means (in feature space) for each image individually to obtain class-agnostic segments, then computes the average feature within each segment, and finally clusters segments over the entire dataset (by using K -means with $K = 21$, as above).

Qualitative results are shown in Fig. 5. Most notably, in contrast to [74], we are able to segment multiple different semantic regions in the same image. For example, in the third row of the figure, our method correctly segments an image containing both a dog and a cat, whereas MaskContrast is limited to a single label (cat) for both regions.

4.4. Image Matting

Finally, we show that with a slight modification, our method can be used for real-world image editing tasks. The only modification we make is that we perform the spectral clustering at full resolution rather than at a reduced intermediate resolution. We sparsify the feature affinity matrix W_{feat} by randomly subsampling $\frac{1}{16^2} \approx 0.4\%$ of its entries. Note that despite the sparsity of W , computing its eigen-

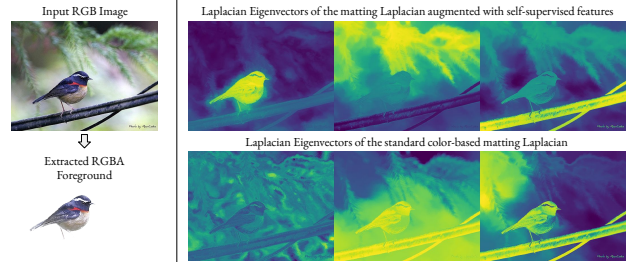


Figure 6. **Image matting.** Combining semantic and color information yields soft image mattes that are well-suited to image editing. Left: the original image and the extracted foreground object using our matting formulation. Right (above): eigensegments derived from our affinity matrix correspond to semantically consistent regions. Right (below): eigensegments from traditional color-based matting do not capture semantic regions well.

values at full ($MN \times MN$) resolution is still too slow to be performed on an entire dataset,² which is why we use a lower resolution for the other experiments. However, if we are only concerned with editing a small number of images, then computing at full resolution is perfectly feasible.

In Fig. 6, we show examples of our mattes, with and without the feature affinity term. We see that the mattes are significantly more useful for editing with our feature affinity term, as they better correspond to objects in the scene. These mattes can then be used as primitives for further computer graphics operations, such as foreground extraction, selective colorization, or background replacement.

5. Conclusions

In this paper we introduced a method for unsupervised localization, segmentation and matting based on spectral graph theory and deep features. Despite the simple formulation, it achieves state-of-the-art unsupervised performance for these tasks. It is interesting to note that this performance is only achieved with features from transformer architectures and not with CNNs, which we attribute to the inherent functionality of self-attention in transformers that aligns well with dense localization tasks. While spectral graph theory has been relegated to a minor role in the age of deep learning, we find that the inductive biases on which it is built can be very useful in the unsupervised setting.

Acknowledgements

L. M. K. is supported by the Rhodes Trust. C. R. is supported by Innovate UK (project 71653) on behalf of UK Research and Innovation (UKRI) and by the Department of Engineering Science at Oxford. I. L. and A. V. are supported by the VisualAI EPSRC programme grant (EP/T028572/1).

²Computing the eigenvectors of W takes around 1 min on 100 CPUs, which corresponds to ≈ 10 days for all of PASCAL VOC 2012.

References

- [1] Rameen Abdal, Peihao Zhu, Niloy Mitra, and Peter Wonka. Labels4free: Unsupervised segmentation using stylegan. In *Proc. ICCV*, 2021. [3](#)
- [2] Yagiz Aksoy, Tae-Hyun Oh, Sylvain Paris, Marc Pollefeys, and Wojciech Matusik. Semantic soft segmentation. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 37(4):72:1–72:13, 2018. [4](#)
- [3] Relja Arandjelović and Andrew Zisserman. Object discovery with a copy-pasting gan. *arXiv preprint arXiv:1905.11369*, 2019. [3](#)
- [4] Sara Atito, Muhammad Awais, and Josef Kittler. Sit: Self-supervised vision transformer. *arXiv preprint arXiv:2104.03602*, 2021. [2](#)
- [5] Yaniv Benny and Lior Wolf. Onegan: Simultaneous unsupervised learning of conditional image generation, foreground segmentation, and fine-grained clustering. *Lecture Notes in Computer Science*, page 514–530, 2020. [7](#)
- [6] Adam Bielski and Paolo Favaro. Emergence of Object Segmentation in Perturbed Generative Models. In *Proc. NeurIPS*, volume 32, 2019. [7](#)
- [7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. [2](#)
- [8] 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 International Conference on Computer Vision (ICCV)*, 2021. [2](#), [6](#)
- [9] Lyndon Chan, Mahdi S. Hosseini, and Konstantinos N. Plataniotis. A comprehensive analysis of weakly-supervised semantic segmentation in different image domains. *International Journal of Computer Vision*, 2020. [1](#), [3](#)
- [10] Jeff Cheeger. A lower bound for the smallest eigenvalue of the laplacian. *Proceedings of the Princeton conference in honor of Professor S. Bochner*, pages 195–200, 1969. [3](#)
- [11] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification, 2021. [2](#)
- [12] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. [7](#)
- [13] Mickaël Chen, Thierry Artières, and Ludovic Denoyer. Unsupervised Object Segmentation by Redrawing. In *Proc. NeurIPS*, volume 32, 2019. [2](#), [3](#), [7](#)
- [14] Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang. Knn matting. *IEEE transactions on pattern analysis and machine intelligence*, 35(9):2175–2188, 2013. [4](#)
- [15] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. [2](#)
- [16] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. [2](#)
- [17] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. [2](#)
- [18] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. [2](#)
- [19] Minsu Cho, Suha Kwak, Cordelia Schmid, and Jean Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1201–1210, 2015. [2](#)
- [20] Subhabrata Choudhury, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Unsupervised part discovery from contrastive reconstruction. *Advances in Neural Information Processing Systems*, 34, 2021. [2](#)
- [21] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting spatial attention design in vision transformers, 2021. [2](#)
- [22] Edo Collins, Radhakrishna Achanta, and Sabine Susstrunk. Deep feature factorization for concept discovery. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 336–352, 2018. [2](#), [3](#)
- [23] 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. [7](#)
- [24] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015. [2](#)
- [25] W. E. Donath and A. J. Hoffman. Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development*, 17(5):420–425, 1973. [3](#)
- [26] Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak mathematical journal*, 23(2):298–305, 1973. [3](#)
- [27] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018. [2](#)
- [28] Ben Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet’s clothing for faster inference, 2021. [2](#)
- [29] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020. [2](#)

- [30] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 2010. 2
- [31] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 447–456, 2015. 5
- [32] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Conference on Computer Vision and Pattern Recognition*, 2019. 2
- [33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7
- [34] Olivier J. Hénaff, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv.cs, abs/1905.09272*, 2019. 2
- [35] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers, 2021. 2
- [36] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7014–7023, 2018. 1, 3
- [37] Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, and Jan Kautz. Scops: Self-supervised co-part segmentation. In *Proc. CVPR*, pages 869–878, 2019. 2
- [38] 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 International Conference on Computer Vision*, pages 7334–7344, 2019. 3
- [39] Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding, 2019. 2
- [40] Phillip Isola, Daniel Zoran, Dilip Krishnan, and Edward H Adelson. Learning visual groups from co-occurrences in space and time. *arXiv preprint arXiv:1511.06811*, 2015. 7
- [41] Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information distillation for unsupervised image segmentation and clustering. *arXiv preprint arXiv:1807.06653*, 2018. 1, 3, 7
- [42] Xu Ji, João F. Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019. 7
- [43] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proc. ICML*, 2021. 1
- [44] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21798–21809. Curran Associates, Inc., 2020. 2
- [45] A. Kanezaki. Unsupervised Image Segmentation by Back-propagation. In *Proc. ICASSP*, pages 1543–1547, 2018. 7
- [46] Tsung-Wei Ke, Jyh-Jing Hwang, and Stella X Yu. Universal weakly supervised segmentation by pixel-to-segment contrastive learning. In *International Conference on Learning Representations*, 2021. 1, 3
- [47] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 3
- [48] Gunhee Kim and Antonio Torralba. Unsupervised detection of regions of interest using iterative link analysis. In *Proc. NeurIPS*, 2009. 6
- [49] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7
- [50] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 7
- [51] Philip Lee and Ying Wu. Nonlocal matting. In *CVPR 2011*, pages 2193–2200. IEEE, 2011. 4
- [52] Junning Li, Pan Zhou, Caiming Xiong, and Steven Hoi. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations*, 2021. 2
- [53] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6
- [54] 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, 2021. 2
- [55] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proc. ECCV*, pages 181–196, 2018. 1
- [56] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Finding an unsupervised image segmenter in each of your deep generative models. *arXiv preprint arXiv:2105.08127*, 2021. 3, 7
- [57] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Proc. NeurIPS*, 2001. 3
- [58] Duc Tam Nguyen, Maximilian Dax, Chaithanya Kumar Mummadi, Thi Phuong Nhung Ngo, Thi Hoai Phuong Nguyen, Zhongyu Lou, and Thomas Brox. DeepUSPS: Deep Robust Unsupervised Saliency Prediction With Self-Supervision. In *Proc. NeurIPS*, 2019. 7

- [59] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016. 2
- [60] Pedro O O. Pinheiro, Amjad Almahairi, Ryan Benmalek, Florian Golemo, and Aaron C Courville. Unsupervised learning of dense visual representations. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4489–4500. Curran Associates, Inc., 2020. 2
- [61] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2
- [62] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 2
- [63] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1
- [64] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000. 3
- [65] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *PAMI*, 22(8), 2000. 4
- [66] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *IEEE transactions on pattern analysis and machine intelligence*, 38(4):717–729, 2015. 7
- [67] Oriane Siméoni, Gilles Puy, Huy V. Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. In *Proc. BMVC*, November 2021. 3, 5, 6, 7
- [68] Yanan Sun, Chi-Keung Tang, and Yu-Wing Tai. Semantic image matting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 4
- [69] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv.cs*, abs/1906.05849, 2019. 2
- [70] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6827–6839. Curran Associates, Inc., 2020. 2
- [71] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers and distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020. 2
- [72] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. *arXiv preprint arXiv:2103.17239*, 2021. 2
- [73] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013. 6
- [74] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. In *International Conference on Computer Vision*, 2021. 1, 3, 7, 8
- [75] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 2
- [76] Huy V Vo, Francis Bach, Minsu Cho, Kai Han, Yann LeCun, Patrick Pérez, and Jean Ponce. Unsupervised image matching and object discovery as optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8287–8296, 2019. 2
- [77] Huy V Vo, Patrick Pérez, and Jean Ponce. Toward unsupervised, multi-object discovery in large-scale image collections. In *European Conference on Computer Vision*, pages 779–795. Springer, 2020. 2, 6
- [78] Huy V. Vo, Patrick Perez, and Jean Ponce. Toward unsupervised, multi-object discovery in large-scale image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 6
- [79] Huy V Vo, Elena Sizikova, Cordelia Schmid, Patrick Pérez, and Jean Ponce. Large-scale unsupervised object discovery. *arXiv preprint arXiv:2106.06650*, 2021. 6
- [80] Andrey Voynov, Stanislav Morozov, and Artem Babenko. Object segmentation without labels with large-scale generative models. In *Proc. ICML*, pages 10596–10606. PMLR, 2021. 3, 7
- [81] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical report, 2011. 7
- [82] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 136–145, 2017. 7
- [83] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021. 2
- [84] Xiu-Shen Wei, Chen-Lin Zhang, Jianxin Wu, Chunhua Shen, and Zhi-Hua Zhou. Unsupervised object discovery and co-localization by deep descriptor transformation. *Pattern Recognition*, 88:113–126, 2019. 6
- [85] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. *arXiv preprint arXiv:2103.15808*, 2021. 2
- [86] Zizhao Wu, Yunhai Wang, Ruyang Shou, Baoquan Chen, and Xinguo Liu. Unsupervised co-segmentation of 3d shapes via affinity aggregation spectral clustering. *Computers and Graphics*, 37(6):628–637, 2013. 3
- [87] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 2
- [88] 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 *Proc. CVPR*, 2021. 2
- [89] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Proc. CVPR*, pages 3166–3173. IEEE, 2013. 7
- [90] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*, 2021. 2
- [91] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12310–12320. PMLR, 2021. 2
- [92] Xiaohang Zhan, Xingang Pan, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised learning via conditional motion propagation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, June 2019. 7
- [93] Runsheng Zhang, Yaping Huang, Mengyang Pu, Jian Zhang, Qingji Guan, Qi Zou, and Haibin Ling. Object discovery from a single unlabeled image by mining frequent itemsets with multi-scale features. *IEEE Transactions on Image Processing*, 29:8606–8621, 2020. 3
- [94] Runsheng Zhang, Yaping Huang, Mengyang Pu, Jian Zhang, Qingji Guan, Qi Zou, and Haibin Ling. Object discovery from a single unlabeled image by mining frequent itemsets with multi-scale features. *IEEE Trans. Image Process.*, 29:8606–8621, 2020. 6
- [95] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. 2, 7
- [96] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *CVPR*, 2017. 2
- [97] Xiao Zhang and Michael Maire. Self-supervised visual representation learning from hierarchical grouping. In *Proc. NeurIPS*, volume 33, 2020. 3
- [98] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Zihang Jiang, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer, 2021. 2
- [99] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European conference on computer vision*, pages 391–405. Springer, 2014. 6