Unsupervised Segmentation by Diffusing, Walking and Cutting

Daniela Ivanova
University of Glasgow,
UK

Marco Aversa
Independent,
Switzerland

Paul Henderson
University of Glasgow,
UK

John Williamson
University of Glasgow,
UK

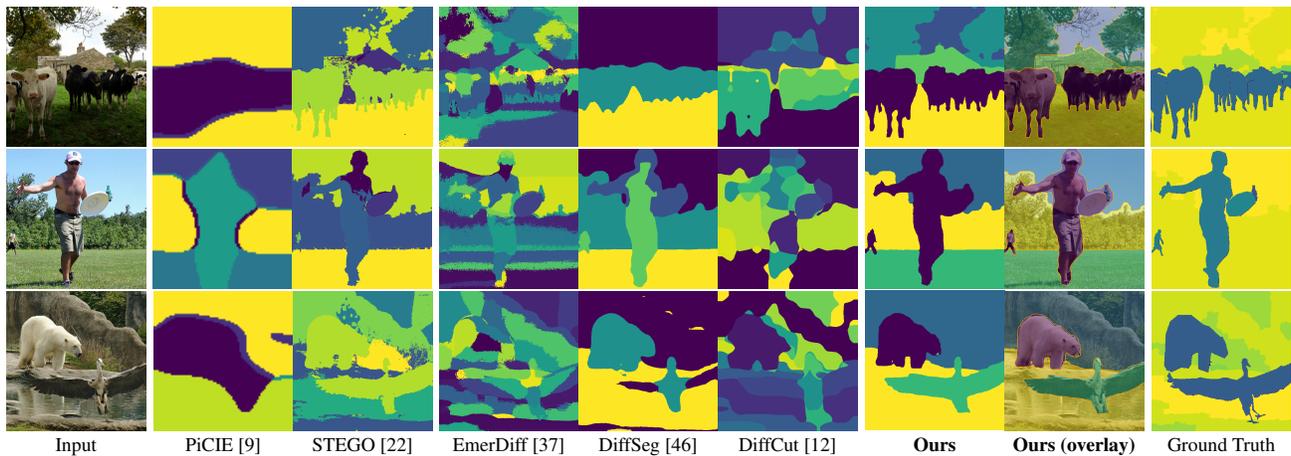| Input | PiCIE [9] | STEGO [22] | EmerDiff [37] | DiffSeg [46] | DiffCut [12] | **Ours** | **Ours (overlay)** | Ground Truth |

Figure 1. We perform unsupervised segmentation by applying Normalised Cuts [42] to self-attention features from Stable Diffusion [41] in a hyperparameter-free setting, and achieve superior performance on Coco-Stuff-27 [5] compared to trained and zero-shot methods.

Abstract

We propose a zero-shot unsupervised image segmentation method by utilising self-attention activations extracted from Stable Diffusion. We demonstrate that self-attention can directly be interpreted as transition probabilities in a Markov random walk between image patches. This property enables us to modulate multi-hop relationships through matrix exponentiation, which captures k-step transitions between patches. We then construct a graph representation based on self-attention feature similarity and apply Normalised Cuts to cluster them. We quantitatively analyse the effects of incorporating multi-node paths when constructing the NCuts adjacency matrix, showing that higher-order transitions enhance hierarchical relationships in the proposed segmentations. Finally, we describe an approach to automatically determine the NCut threshold criterion, avoiding the need to manually tune it. Our approach surpasses all existing methods for zero-shot unsupervised segmentation based on pretrained diffusion models features, achieving state-of-the-art results on COCO-Stuff-27, Cityscapes and ADE20K.

1. Introduction

Segmentation of images into semantically-meaningful regions is of key interest in computer vision. *Unsupervised semantic segmentation* aims to label images without explicit training to identify specific object types [22, 25, 42]. Zero-shot segmentation eschews training altogether, repurposing models which have not been trained for a segmentation task at all, with the goal of improved generalization [12, 27, 45, 46]. We focus on *unsupervised zero-shot multi-class semantic segmentation*, where an image is labelled with many distinct and interpretable labels ("horse, sand, human, sky") rather than e.g. binary partitioning into foreground/background as in [27, 35, 51].

Modern approaches to unsupervised segmentation leverage the information inherently captured in the latent features of models trained for other tasks. While some methods require access to the target data for downstream unsupervised segmentation training [6, 21, 22, 28, 39, 43], and hence, cannot segment outside of their target domain, zero-shot approaches solve this by directly utilizing the features of generative text-to-image models like Stable Diffusion [41] for semantic [12, 45, 46] or even interactive open-vocabulary segmentation [27]. Stable Diffusion's

self-attention features richly encode how much each (latent) pixel attends to every other pixel, and form a natural basis for partitioning an image [46]. Because the self-supervised training process needs to parsimoniously account for e.g. spatial equivariance of rigid objects or deformation of flexible objects like human bodies, the latent features would be expected to give rise to similar embeddings for pixels across objects that are consistent under these types of transformations.

Leveraging the effectiveness of self-attention features from Stable Diffusion 1.4, as in DiffSeg [46], we propose a novel unsupervised zero-shot segmentation approach. Departing from DiffSeg's bottom-up, iterative merging strategy, we employ the classic Normalised Cuts algorithm [42] to iteratively partition image regions based on self-attention similarity. To address the challenge of manual tuning in Normalised Cuts, we propose a method that automatically determines the stopping criterion. This yields a tuning-free method that automatically estimates optimal thresholds for halting the iterative partitioning, surpassing the performance of existing zero-shot unsupervised segmentation techniques.

Furthermore, we interpret the self-attention maps as transition kernels for a random walk across the image space, effectively grouping latent pixels with high connectivity probability. By exponentiating these kernels, we enable random walks to take larger steps, providing a natural and controllable mechanism for selecting the desired semantic granularity of the segmentation. This mathematical interpretation not only provides theoretical grounding for our approach but also enables intuitive control over segmentation detail.

Our contributions are:

- A zero-shot unsupervised semantic segmentation algorithm that outperforms the state-of-the-art, including DiffSeg [46], EmerDiff [37], and DiffCut [12];
- A random walk interpretation of self-attention features used to build the Normalised Cuts adjacency matrix, which can modulate semantic granularity by simple matrix exponentiation;
- A novel approach to *automatically* selecting the critical threshold for Normalised Cuts on a per-image basis that eliminates the main "fiddle factor" in unsupervised segmentation.

2. Related Work

2.1. Unsupervised Segmentation via Generative Models

Internal features of GANs [19], VAEs [29] and other self-supervised models such as DINO [1, 6], have long been used for diverse downstream image tasks, achieving impressive results [7, 10, 22, 31, 32, 47, 48]. Recently, large-scale generative text-to-image models (e.g. Stable Diffu-

sion [41]), have provided even richer internal features to be used for downstream tasks including depth estimation [8], semantic correspondence [23, 33, 45, 53], and segmentation [3, 52]. Unsupervised zero-shot segmentation is one example where methods leveraging diffusion model features have achieved state-of-the-art results [12, 37, 46]. **EmerDiff** [37] demonstrates that cross-attention query vectors encode semantic information at each spatial location. They first perform k-means clustering on these vectors at low resolution to obtain initial segments. To achieve high-resolution segmentation, they perturb the cross-attention activations corresponding to each segment, re-generate, and observe which pixels in the output are most affected. Although this method produces accurate segmentation boundaries, it requires two denoising passes through the diffusion model *per segment*. Additionally, the approach requires prior knowledge of the number of segments, meaning it is not truly unsupervised.

DiffSeg [46] instead uses self-attention maps from Stable Diffusion [41]. For an input image yielding N latent patches, they extract an N×N self-attention matrix where each row represents the attention distribution from one patch over all other patches. Inspired by SAM [30], they sample "anchor" points on a grid and iteratively merge regions based on the Kullback-Leibler divergence between their averaged attention distributions. This bottom-up approach does not require the number of segments to be provided *a priori*. However, it is computationally expensive due to both the grid sampling process and KL divergence calculations between large matrices; the latter must be calculated twice between each pair of patches/segments as it is not symmetric. Furthermore, the grid sampling strategy may miss small but significant image features if the anchors aren't sampled densely enough, and the iterative merging requires pre-setting a threshold over KLD as well the number of iterative merging passes.

DiffCut [12], inspired by DiffSeg, extracts features (rather than raw self-attentions) from a Stable Diffusion XL model, and employs Normalised Cuts [42] iteratively with a manually tuned stopping criterion, and an exponentiation factor to increase the contrast of the adjacency matrix for easier partitioning. This follows from a line of recent works which has revitalized the use of spectral clustering methods by leveraging graphs constructed from neural network features. Deep Spectral Methods [35], for instance, utilize features extracted from a pre-trained self-supervised vision transformer (DINO) for salient object localization. Similarly, TokenCut [51] and MaskCut [50] employ Normalised Cuts on DINO features, with MaskCut iteratively partitioning the image into a pre-defined number of segments. More recently, Panoptic Cut [26] extends this approach by applying Normalised Cuts for a fixed number of iterations on DINO features, generating segment proposals for open

vocabulary segmentation. Concurrent works interpret vision transformer attention activations as transition probabilities [15, 17, 27]. Notably, [54] first explored the effectiveness of spectral clustering over features from generative diffusion models, demonstrating they encode rich semantic and spatial information.

A common limitation across all the aforementioned segmentation methods is their reliance on either a single bipartition for salient object detection (Deep Spectral Methods, TokenCut) or iterative partitioning with manually specified parameters like the number of clusters (MaskCut), partitioning iterations (PanopticCut) or stopping criteria (DiffCut). Instead, our approach extracts and aggregates self-attention activations from Stable Diffusion 1.4 as in DiffSeg, to which we apply Normalised Cuts, concurrently to DiffCut. In addition, we propose a way to automatically halt the partitioning of each image, eliminating the need for manual tuning of NCut parameters, such as number of iterations or stopping threshold. Finally, we observe a special property of self-attention activations in particular, which allows us to control the partitioning for semantic granularity. Next, we review the theory behind the classic Normalised Cuts algorithm which informed our approach.

2.2. Normalised cuts

Spectral graph theory traces its roots to the 1970s [14, 18], but was first famously applied to computer vision by Shi and Malik [42] with the advent of Normalised Cuts.

Here, an image is represented as a weighted, undirected graph $G = (V, E, w)$ where pixels correspond to vertices V, and edge E weights w represent some similarity measure between each pair of pixels. The graph can be partitioned, or cut, into two disjoint sets A and B, where the cost of cutting it is the sum of all edge weights that are removed:

$$\text{cut}(A, B) = \sum_{u \in A, v \in B} w(u, v). \tag{1}$$

The normalised cut (NCut) instead normalizes this cost relative to the total connections in each part of the graph:

$$\text{NCut}(A, B) = \frac{\text{cut}(A, B)}{\text{assoc}(A, V)} + \frac{\text{cut}(A, B)}{\text{assoc}(B, V)} \tag{2}$$

Here $\text{assoc}(A, V)$ measures the total connection (weighted sum of all edges) from nodes in A to all nodes. This normalization prevents the algorithm from isolating individual vertices, since cutting off a small set would result in a high NCut value. Instead the minimum-cost cut tends to split the image into two large but dissimilar regions.

The authors show that the normalised cut formulation can be solved through a generalized eigenvalue problem. Given a adjacency matrix A (representing the weighted graph G) containing edge weights between nodes and its

corresponding degree matrix D (where each diagonal entry is the sum of that row in A), the NCut optimisation is equivalent to solving:

$$Lx = \lambda D x, \tag{3}$$

where L is the Laplacian $L = D - A$. The solution is found in the eigenvector x_2 corresponding to the second-smallest eigenvalue λ_2, also known as the Fiedler vector, which is used to describe a graph's algebraic connectivity [18]. This eigenvector assigns a value to each pixel; the algorithm finds a threshold in these values that minimizes the Normalised Cut, bipartitioning the image. Since the Fiedler vector takes on continuous values, the authors recommend testing several thresholds, and selecting the one with the smallest NCut cost. They also extend the algorithm to give more than two regions, by iteratively partitioning each resulting subgraphs until some condition, such as a target number of clusters, or a threshold NCut cost, is met.

Dynamic NCut threshold. A further extension to the NCuts algorithm is the Minimum Average Node Cut (MAN-C) [49], which aims to overcome the need for manually setting the number of clusters, iterations, or threshold NCut cost. It defines a "good" cluster based on the condition min-cut$(G) \geq \frac{nT}{2m}$, where n is the number of vertices, T is the total edge weight, and m is the number of edges. In MAN-C, the *minimum cut* for a partition is the smallest sum of edge weights needed to separate the graph into two parts. For each proposed Normalised Cuts partition, MAN-C checks if this minimum cut is lower than the threshold $\frac{nT}{2m}$ to determine when to halt the recursive partitioning.

Random walk interpretation. Meila and Shi [34] later gave a probabilistic interpretation of NCut by showing that spectral clustering can be viewed through the lens of random walks on graphs. They demonstrated that the degree-normalised adjacency matrix $P = D^{-1}A$ represents transition probabilities of a random walk, where $P_{i,j}$ represents the probability of moving from node i to node j. The eigenvectors of this transition matrix correspond exactly to the NCut solution, with the second largest eigenvector providing the optimal binary partition. This interpretation reveals that spectral clustering effectively groups nodes that have high probability of being connected by random walks and suggests that any matrix with random walk properties (row-stochastic, capturing similarity relationships) can be directly used in the NCut framework.

3. Methods

The success of DiffSeg [46] and EmerDiff [45] demonstrates that pre-trained text-to-image diffusion models encode rich semantic relationships in their internal representations. Our approach recognizes and exploits a fundamental mathematical property of these models' self-attention matrices. In this section, we present our method in five parts:

first, we describe how we extract self-attention features from Stable Diffusion (Section 3.1); next, we explore two complementary approaches to leveraging these features—directly interpreting self-attention as random walks (Section 3.2) and constructing adjacency matrices from attention feature similarity (Section 3.3); we then detail our application of Normalized Cuts to these representations (Section 3.4); propose an automatic threshold determination method (Section 3.5); and finally describe our upsampling process (Section 3.6).

3.1. Extracting diffusion features

Given the pre-trained diffusion model, we use DDIM Inversion [13, 44] to invert real images for a small number of steps (we use 10 out of 50 in our experiments, and ablate this in the Supplementary), then extract self-attention from the denoising U-Net decoder from a single generation step. For Stable Diffusion, transformer activations exist in four resolutions across the decoder: 8^2, 16^2, 32^2 and 64^2. We extract self-attention maps $S^{(r)} \in \mathbb{R}^{N \times N}$ where $r \in \{0, 1, 2, 3\}$ is the resolution level and N corresponds to the spatially flattened latent feature resolution 8^2, 16^2, 32^2 and 64^2. We sum across attention heads following [46, 54]. We aggregate over resolutions by upsampling to $64^2 \times 64^2$ followed by a proportional weighted sum, empirically defined in DiffSeg [46], producing an aggregated self-attention matrix $P = \sum_r w_r \cdot S^{(r)}$. Each matrix $S^{(r)}$ is row-normalized through softmax, and since the weights w_r sum to one, the aggregated matrix P maintains row-normalization.

3.2. Self-attention defines a random walk

While traditional NCut requires constructing an adjacency matrix A, from which the random-walk interpretation derives a degree-normalized (i.e. row-stochastic) transition matrix $P = D^{-1}A$, we observe that the aggregated self-attention P already satisfies the mathematical properties of a random walk transition matrix, albeit without being derived explicitly from some adjacency matrix A.

As described in the previous section, our aggregated self-attention matrix maintains row-normalization through the softmax operation in the attention mechanism and weighted aggregation. Consequently, each row in P represents a probability distribution, with elements $P_{i,j}$ giving the probability of transitioning from patch i to patch j in a random walk across the image.

It is therefore possible to directly apply Normalized Cuts theory in the Random Walk reframing. We use the transition matrix as the input to recursive NCut as proposed by Shi and Malik [42], as described in Section 3.4, by adapting the algorithm to use the eigenvectors of the transition probabilities P (c.f. [34]) instead of an adjacency matrix A. Since P is not necessarily symmetric, to estimate the second

eigenvector, we apply power iteration with one deflation.

This formulation provides a principled approach to leveraging self-attention directly for segmentation, without requiring the construction of an intermediate adjacency matrix or similarity metric, while still benefiting from the theoretical guarantees of spectral clustering.

3.3. Adjacency matrices from feature similarity

Beyond directly using self-attention as transition probabilities, we can construct traditional adjacency matrices by treating each patch's attention pattern as a feature embedding. While DiffSeg uses Kullback-Leibler divergence between attention distributions, we employ either dot product or cosine similarity (empirically contrasted in Section 4.2), which can be computed more efficiently.

These constructions yield a positive-valued, symmetric adjacency matrix A suitable for the conventional NCut formulation. This allows for efficient direct eigendecomposition rather than power iteration.

At the same time, we can exploit the random-walk property of the self-attention matrix established in Section 3.2 prior to constructing the adjacency matrix. By computing P^k through matrix exponentiation, each entry $(P^k)_{ij}$ represents the probability of transitioning from patch i to j in exactly k steps [38]. This k-step transition probability aggregates all possible walks of length k between patches i and j, capturing information flow through intermediate patches.

The exponent k serves as a scale parameter: smaller values preserve local attention patterns, while larger values enable broader information diffusion across the image, connecting semantically related but spatially distant patches (see first row of Figure 6). We evaluate several values of k in Section 4.4, showing how this parameter effectively controls the hierarchical level of segmentation.

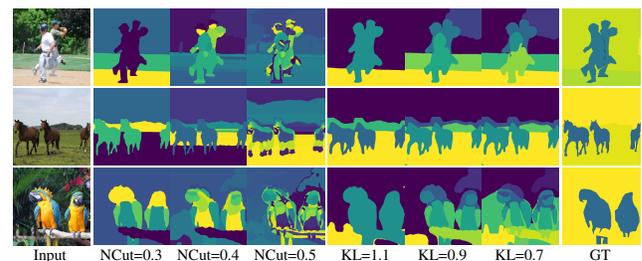| Input | NCut=0.3 | NCut=0.4 | NCut=0.5 | KL=1.1 | KL=0.9 | KL=0.7 | GT |

Figure 2. Qualitative comparison on COCO-Stuff-27 between our simple Random Walk approach for different NCut cost thresholds (columns 2-4) and DiffSeg for different Kullback-Leibler Divergence thresholds (columns 5-7).

3.4. Applying normalised cuts (NCut)

So far, we have introduced two approaches to quantify semantic correlations across the image by (1) modeling the aggregated self-attention matrix as a k-step random walk

among patches via matrix exponentiation, and (2) constructing an adjacency matrix from patch-level feature similarities. As discussed in Section 2.2, both introduced formulations allow us to apply Shi and Malik's recursive Normalised Cuts (NCut) criterion [42] to partition the image. Regardless of whether a random walk or adjacency matrix is used, the algorithm proceeds as follows. (1) Solve the generalised eigenvalue problem (equation 3). (2) Threshold the Fiedler vector across test values, selecting the bipartition with lowest NCut value. (3) If the NCut cost exceeds a predefined threshold, return the current partitions. (4) Otherwise, recursively apply the algorithm to each newly formed partition.

3.5. Adapting MAN-C

A key challenge in Normalized Cuts is determining when to stop recursive partitioning. Manual threshold selection is burdensome and image-dependent. To address this, we adapt the Minimum Average Node Cut (MAN-C) algorithm [49], detailed in Section 2. MAN-C works well for sparse graphs, where there are few edges. However, our graphs derived from self-attention similarities have very high edge density, and in this case the MAN-C threshold reduces to $\frac{T}{n-1}$ i.e. it is scaled by the number of nodes—therefore, a minimum cut will almost always be larger, since it's not appropriately scaled w.r.t. the threshold anymore.

Hence, we propose and evaluate (Section 3.5) two alternative adjustments:

1. Directly compare NCut cost to the MAN-C threshold;
2. Compare to a *scaled* minimum cut, $\frac{\text{min-cut}(G)}{T}$, which scales the minimum cut relative to the graph's total edge weight T, to place it in the same scale as the MAN-C threshold.

Both approaches are designed to keep the adaptive nature of MAN-C while handling the dense nature of our similarity graphs.

3.6. Segmentation aggregation and upsampling

After recursive partitioning, we aggregate the NCut partitions into unique segments and calculate the average feature vector for each region, resulting in K averaged PDFs for K segments of shape $N \times N$, where N is the latent size (number of latent patches). We upsample these segmentations to the original image resolution by first applying bilinear interpolation to the original features—$512 \times 512 \times N^2$, then computing cosine similarity between each upsampled pixel feature and each segment's average distribution. Each pixel is assigned to its closest segment, producing the final segmentation map.

4. Experiments

Datasets. For our main experiments, we follow [9, 25, 37, 46] and use a subset [25] of COCO-Stuff-27 [5]

validation data. It consists of 2174 images, combining 80 thing and 91 stuff categories from COCO-Stuff-171 into 27 classes. Our final approach is also evaluated on Cityscapes [11], following DiffSeg [46]. We additionally evaluate on ADE20K [55], and use it for our comparison with DiffCut [12].

Metrics. To quantitatively evaluate unsupervised segmentation, we report Mean Pixel Accuracy, Mean Intersection over Union (mIoU), as done in previous works [22, 37, 46], in addition to F1 score. Since class labels are not predicted, we use Hungarian matching to assign predicted segments to their closest Ground Truth segments, following standard practice [37, 43, 46]. This penalizes oversegmentation, since when there are more predicted segments than Ground Truth ones, the remaining predicted segments are considered false positives. We note that due to the varying granularity of human annotated segmentation labels, this standard evaluation strategy is far from perfect. For example, in Figure 2, row 1, the Ground Truth merges the human figures, along with attributes such as gloves and helmets into a single annotation. Our method can match this annotation granularity with a lower NCut threshold (row 1, column 2), or separate the baseball glove, along with the hats, shoes, trousers, shirts, legs and arms into their own segments when the threshold is increased. We dicuss how this affects quantitative evaluation in the Supplementary, Section A.

4.1. Random Walk Baseline

As outlined in Section 3.2, our simple baseline approach uses the aggregated self-attention matrices directly as a random walk transition matrix and applies recursive Normalized Cuts over it. We show examples from several thresholds over the NCut cost as a stopping criterion in Figure 2, where we also compare it to results from DiffSeg for their best-perfoming KLD threshold (1.1, column 5) along with two lower values. Since Normalised Cuts is inherently a top-down segmentation approach, we note how a higher NCut threshold increases granularity in the predicted segments, which generally follow a hierarchical structure; for DiffSeg, lower KLD threshold has a similar effect, as the approach merges segments in a bottom-up fashion. Overall, our approach's predictions are more precise, and demonstrate a clear hierarchical semantic relationship between segments across thresholds, even in the case of extreme oversegmention—for instance, the horses legs, hooves, backs and faces are all grouped together in row 2, column 4. DiffSeg predictions, on the other hand, suffer from pronounced horizontal banding artefacts (row 2, 5). Quantitative results for NCut threshold set to 0.3 in Table 4 show that this simple approach already performs almost as well as the state-of-the-art DiffSeg, particularly with respect to mIoU and accuracy.
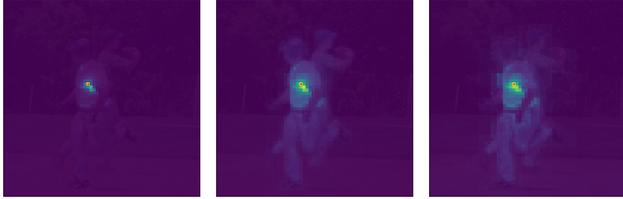
(a) Self-attention PDF (aggregated)

(b) Dot product adjacency

(c) Cosine similarity adjacency

Figure 3. Self-attention PDF (aggregated) (a) and corresponding dot product (b) and cosine similarity (c) adjacency values for a single patch.

4.2. Self-Attention adjacency

The previous experiment motivates a natural follow-up: how can we best construct a traditional NCuts adjacency matrix based on the similarity between self-attention activations? Here, we validate the two variants proposed in Section 3.3: constructing A as the dot-product between the self-attention PDFs, or as their cosine similarity (normalised dot product). Both variants result in a positive-valued symmetric adjacency matrix A, since self-attention is already probability-normalised. We compare the results of applying Normalised Cuts over each variant across three different NCut Thresholds in Table 1, where the Dot Product variant is the clear leader—note that when the NCut threshold is set to 0.4, it already outperforms state-of-the-art DiffSeg in terms of both mIoU and Accuracy.

Examining the spatial distribution plots of self-attention compared to the two adjacency variants for a given patch marked by red dot in Figure 3, we can see the intuition behind why dot product performs better. For a given query position q, the self-attention produces a PDF p_q that is sharply peaked at q and its immediate neighborhood $N(q)$, with near-zero values elsewhere, as shown in (a). The dot product $p_q^T p_r$ between two such distributions preserves this sparsity structure, showing high similarity only when neighborhoods overlap in (b). In contrast, cosine similarity $\frac{p_q^T p_r}{||p_q||||p_r||}$ normalizes away this locality-preserving property, artificially inflating similarities between distributions with disjoint support, as evidenced by more distributed similarities in (c), which form a halo around the object of interest. Hence, more uniformly distributed adjacency values would have a negative effect on Normalised Cuts, as it relies on having strong intra-segment connections and weak inter-segment connections; cosine similarity would therefore make the distinction between segments less pronounced and lead to arbitrary cuts, as seen in Figure 4.

4.3. Hyperparameter-free NCut

In Section 3.5 we proposed a way to perform Normalised cuts without setting an NCut cost as a stopping criterion, making the approach completely hyperparameter-free.

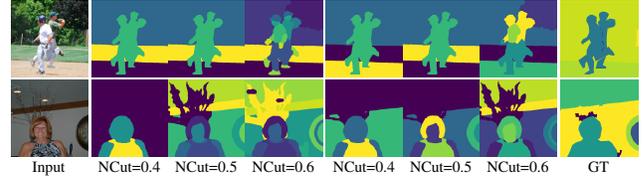Input NCut=0.4 NCut=0.5 NCut=0.6 NCut=0.4 NCut=0.5 NCut=0.6 GT

Figure 4. Qualitative comparison on COCO-Stuff-27 between NCut over a dot-product Adjacency matrix (columns 2-4) and over a cosine similarity Adjacency matrix (columns 5-7) across a range of NCut thresholds.

Model Variant	NCut Threshold	Acc ↑	F1 ↑	mIoU ↑
Dot Product	0.4	74.0	58.4	44.0
	0.5	71.6	56.7	43.1
	0.6	60.0	45.2	32.1
Cosine Similarity	0.4	72.4	56.0	41.7
	0.5	71.5	56.0	42.2
	0.6	62.9	47.7	34.4

Table 1. Quantitative results on COCO-Stuff-27 between Normalised Cuts performed on adjacency matrices build as dot product vs cosine distance between patch self-attention across three different NCut thresholds.

Specifically, we suggested two versions of the MAN-C algorithm with thresholds adapted to the specific graph structure that arises from our adjacency matrices. We compare these two automatic thresholding approaches to assess their effectiveness in partitioning the self-attention adjacency graphs, using the dot product construction from the previous experiments. While the quantitative results in Table 3 demonstrate that comparing the MAN-C criterion with NCuts is much more effective, we can see in the qualitative examples in Figure 5 that the segmentations proposed by thresholding with the scaled MinCut are not necessarily bad, as semantically meaningful groupings are clearly emerging; rather, they are too granular, and punished by the Hungarian matching due to the coarse annotations, as discussed earlier. Additionally, since the approach is now free of hyperparameters, we follow up with an ablation study on the self-attention resolution levels $8, 16, 32, 64$ by progressively removing each, starting from the lowest resolution. This allows us to evaluate how different self-attention scales contribute to segmentation quality. Removing lower resolutions increases the granularity of predictions for both methods, as seen in Figure 5, and decreases the metrics they achieve; yet the qualitative examples demonstrate preserved semantic coherence even for the most granular predictions, such as those made with the Scaled MinCut threshold. Overall, our quantitative results show that we outperform DiffSeg, which optimises several dataset-specific hyperparameters (merging level, number of anchors, and KLD merging threshold) in a completely hyperparameter-free setting.
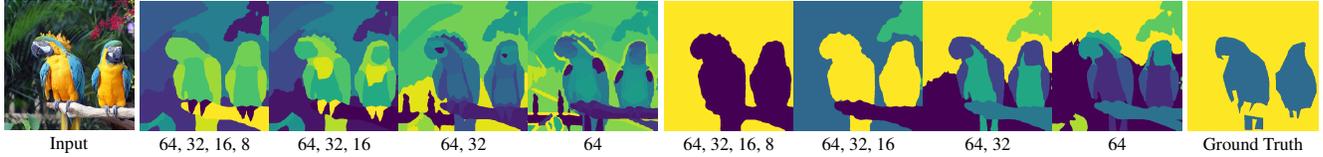
Figure 5. Qualitative comparison of segmentation results for our two automatic thresholding approaches: Scaled MinCut (columns 2-5) and NCut (columns 6-9) across different self-attention resolution levels.

steps k	1	2	3	4	5
Acc ↑	45.9	67.0	71.2	70.7	70.0
F1 ↑	36.1	54.4	56.9	55.8	54.2
mIoU ↑	23.8	40.7	42.4	41.1	39.6

Table 2. Comparison for a range exponents (random walk steps k) applied to the self-attention transition matrix P before constructing the adjacency matrix A. We employ Scaled MinCut for automatic threshold selection, considering all available resolutions.

4.4. Power Walks

While our previous experiment demonstrates a fully automated approach for zero-shot unsupervised segmentation, removing the need for pre-setting the NCut threshold means we do not have a way to control the granularity of segmentations. As established in Section 3.3, self-attention matrices can be interpreted as random walk transition probabilities. This allows us to modulate multi-hop relationships through matrix exponentiation prior to constructing our adjacency matrix. Our quantitative results in Table 2 show that considering higher order walks ($k = 2, 3, 4$) substantially improves the performance of the Scaled MinCut threshold approach against the annotations of COCO-Stuff-27. Qualitatively (Figure 6), we see that higher exponents have the effect of hierarchically merging the more granular segments produced at $k = 1$: the heads, wings, feet and bodies of the parrots are merged into a larger "parrot" segment, which is closer to the human-annotated ground truth labels.

4.5. Comparison with state-of-the-art

In Table 4 we summarise the results of our method alongside several recent approaches to unsupervised segmentation—DiffSeg [46], DiffCut [12] and EmerDiff [37], which are the closest to it conceptually, being zero-shot unsupervised segmentation models which rely on Stable Diffusion features and represent state-of-the-art. We also report results from *trained* unsupervised state-of-the-art methods PiCIE [9], STEGO [22], DepthG [43], U2Seg [39], CUPS [21] and EAGLE [28].

For DiffSeg, we use their suggested best hyperparameters, i.e. 1.1 and 0.9 KLD threshold for COCO-Stuff and Cityscapes respectively, and report additional ADE20K results from [12]. For EmerDiff, we evaluate with Hungarian matching (dubbed "traditional evaluation strategy" by the authors) for a fair comparison against the other meth-

Approach	Resolution Levels				Acc ↑	F1 ↑	mIoU ↑
	8	**16**	**32**	**64**			
Random Walk	✓	✓	✓	✓	72.2	56.9	43.3
MAN-C w/ NCut	✓	✓	✓	✓	74.7	60.6	<u>46.1</u>
		✓	✓	✓	74.1	60.8	**46.6**
			✓	✓	70.1	56.4	42.7
				✓	63.9	49.6	36.4
MAN-C w/ Scaled MinCut	✓	✓	✓	✓	45.9	36.1	23.8
		✓	✓	✓	42.0	32.0	20.6
			✓	✓	37.2	27.6	17.0
				✓	36.1	25.0	15.3
DiffSeg	✓	✓	✓	✓	72.5	58.5	43.6
DiffCut			✓		59.6	44.5	31.3

Table 3. Comparison of our automatic threshold methods across different attention resolution combinations on Coco-Stuff-27. MAN-C NCut consistently outperforms both Scaled MinCut and state-of-the-art DiffSeg. Random Walk baseline has NCut threshold set to 0.3.
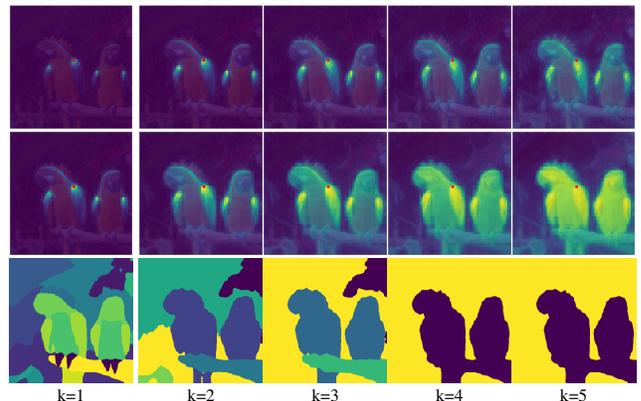

Figure 6. Power Walks: Self-attention extracted from Stable Diffusion for an arbitrary patch marked by red dot (row 1) framed as a random walk transition probability matrix and raised to the kth power, allows us to build adjacency matrices which increasingly capture higher-order patch-similarities (row 2). The resulting unsupervised segmentations (row 3) produced by our hyperparameter-free NCut method demonstrate that as long-range relationships are enhanced, the granularity of segmentation decreases, while still extracting semantically-meaningful segments.

	Trained							**Zero-shot**		
	PiCIE [9]	**STEGO** [22]	**DepthG** [43]	**U2Seg** [39]	**CUPS** [21]	**EAGLE** [28]	**EmerDiff** [37]	**DiffSeg** [46]	**DiffCut** [12]	**Ours**
COCO-Stuff	13.8	28.2	29.0	30.2	-	27.2	26.6	<u>43.6</u>	31.3	**46.6**
Cityscapes	12.3	21.0	23.1	21.6	26.8	22.1	12.0	<u>21.2</u>	18.1	**22.0**
ADE20K	-	-	-	-	-	-	-	37.7	<u>40.1</u>	**41.2**

Table 4. Quantitative comparison of mIoU between our approach and current unsupervised zero-shot state-of-the-art on COCO-Stuff-27, Cityscapes and ADE20K, along with non-zero-shot unsupervised state-of-the-art for completeness. Best in bold, second best underscored.

ods, and set the mask proposal parameter $K = 30$, as the authors suggest. While both DiffSeg and EmerDiff rely on Stable Diffusion 1.4 as a feature extractor, DiffCut report their main results on SSD 1B [20], a distilled variant of Stable Diffusion XL[40], and utilise a post-processing step to refine segmentation predictions - PAMR [2]. For fairness, we exclude PAMR, and evaluate the method on Stable Diffusion 1.4 using hyperparameters $\tau = 0.2$ and $\alpha = 10$, optimised for SD1.4 by the authors in their ablation. We include the best-performing variant of our method, which uses three self-attention resolutions, and moreover does not rely on any hyperparameters as it leverages our dynamic NCut thresholding scheme; this variant improves on the state-of-the-art DiffSeg by 3.0% mIoU on COCO-Stuff and 0.8% on Cityscapes. In fact, some of the weaker variants of our approach still outperform DiffSeg, see Tables 1 & 3. Furthermore, our dynamic threshold approach outperforms DiffCut, despite it using NCut hyperparameters specifically optimised for Stable Diffusion 1.4. In terms of computational cost, all variants take 2-4s per image, with the highest-scoring one (w.r.t. quantitative metrics) taking 2.5s on an A5000 GPU and using 16GB of VRAM.

Model	Method	w/o PAMR	w/ PAMR
SSD 1B	DiffCut	46.1	49.1
	Ours	41.3	42.6
SD 1.4	DiffCut	31.3	45.2
	Ours	46.6	47.1

Table 5. Comparison with DiffCut (mIoU) on COCO-Stuff-27.

We perform further ablation against DiffCut Table 5. For fair comparison, we benchmark our approach against Diff-Cut using both SD 1.4 and SSD 1B. We also ablate the effect PAMR postprocessing by adding it to our pipeline as well. We use the best hyperparameters suggested by the authors, i.e. $\alpha = 10$ and $\tau = 0.5$ for SSD 1B or $\tau = 0.2$ for SD 1.4.

With SD 1.4, our method achieves 46.6 mIoU on Coco-Stuff-27 without PAMR post-processing outperforming DiffCut's 45.2 mIoU *with* PAMR and 31.3 without. In fact, without PAMR, our SD 1.4 results (46.6 mIoU) outperform DiffCut even with the newer, larger SSD 1B model.

Unlike SD 1.4, SSD 1B lacks self-attention at 8×8 and 16×16 resolutions, and has an asymmetric encoder-decoder structure with additional decoder transformer blocks (34 in total), hence the DiffSeg feature aggregation strategy does not translate to this setting. While our initial results with this backbone show promise (Table 5), optimization of feature aggregation strategies for SSD 1B might improve performance.

Qualitatively, both our coarser and finer predictions (Figure 1) are more precise than those of DiffSeg and DiffCut, with additional visual comparisons in the Supplementary.

5. Conclusions

Unsupervised image segmentation is a challenging and ambiguous task where even the definition of "correct" remains debatable [16]. We demonstrate that we can exploit the naturally emerging patterns in diffusion self-attentions to discover meaningful segmentations, by exploring the activations through the lens of random walks. By framing self-attention as transition probabilities in a Markov process, we establish a principled approach for clustering image regions based on their semantic relationships, while offering intuitive control over segmentation granularity through walk length. Our approach combines the strengths of both top-down spectral clustering methods (like DiffCut) and bottom-up merging strategies (like DiffSeg) while addressing their limitations: we eliminate the need for manual threshold tuning in recursive partitioning through our automatic threshold determination, while avoiding the grid sampling and iterative merging inefficiencies of bottom-up approaches through our power walk exponentiation, which naturally handles the "merging" of semantically related regions. This balanced methodology produces hierarchically consistent segmentations that align with human perceptual judgments across diverse image types.

Our approach reveals intrinsic object groupings in learned representations, specifically self-attention layers, while achieving superior quantitative results. Though implemented for Stable Diffusion 1.4, our methodology is model agnostic and applicable to any model with attention features, as demonstrated with SSD 1B in our ablation against DiffCut.

Variants that most closely matched human annotation granularity naturally produced the best quantitative results, while qualitative analysis reveals that even lower-scoring variants maintain consistent semantic groupings. Future work should investigate the discrepancies between human annotations and features learned by foundation models.

Acknowledgements

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep ViT features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2021. 2

[2] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4253–4262, 2020. 8

[3] Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *International Conference on Learning Representations*, 2021. 2

[4] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad – a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 4

[5] 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. 1, 5, 2

[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. 1, 2

[7] Jiaqi Chen, Jiachen Lu, Xiatian Zhu, and Li Zhang. Generative semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7111–7120, 2023. 2

[8] Yida Chen, Fernanda Viégas, and Martin Wattenberg. Beyond surface statistics: Scene representations in a latent diffusion model, 2023. *NeurIPS 2023 Workshop on Diffusion Models*, 2023. 2

[9] 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. 5, 7, 8, 1

[10] Edo Collins, Raja Bala, Bob Price, and Sabine Susstrunk. Editing in style: Uncovering the local semantics of gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5771–5780, 2020. 2

[11] 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 *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5, 1, 3

[12] Paul Couairon, Mustafa Shukor, Jean-Emmanuel Haugeard, Matthieu Cord, and Nicolas Thome. Diffcut: Catalyzing zero-shot semantic segmentation with diffusion features and recursive normalized cut. *Advances in Neural Information Processing Systems*, 37:13548–13578, 2024. 1, 2, 5, 7, 8

[13] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 4

[14] William E Donath and Alan J Hoffman. Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development*, 17(5):420–425, 1973. 3

[15] Batu El, Deepro Choudhury, Pietro Liò, and Chaitanya K Joshi. Towards mechanistic interpretability of graph transformers via attention graphs. *arXiv preprint arXiv:2502.12352*, 2025. 3

[16] Mark Eramian, Christopher Power, Stephen Rau, and Pulkit Khandelwal. Benchmarking human performance in semi-automated image segmentation. *Interacting with Computers*, 32(3):233–245, 2020. 8

[17] Yotam Erel, Olaf Dünkel, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Amit H Bermano. Attention (as discrete-time markov) chains. *arXiv preprint arXiv:2507.17657*, 2025. 3

[18] Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak mathematical journal*, 23(2):298–305, 1973. 3

[19] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, page 2672–2680, Cambridge, MA, USA, 2014. MIT Press. 2

[20] Yatharth Gupta, Vishnu V. Jaddipal, Harish Prabhala, Sayak Paul, and Patrick Von Platen. Progressive knowledge distillation of stable diffusion xl using layer level loss, 2024. 8

[21] Oliver Hahn, Christoph Reich, Nikita Araslanov, Daniel Cremers, Christian Rupprecht, and Stefan Roth. Scene-centric unsupervised panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24485–24495, 2025. 1, 7, 8

[22] 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, 5, 7, 8

[23] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[24] Daniela Ivanova, Marco Aversa, Paul Henderson, and John Williamson. State-of-the-art fails in the art of damage detection. *European Conference on Computer Vision (ECCV) Workshop on VISART, 2024*, 2024. 1, 4

[25] 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, 5

[26] Dahyun Kang and Minsu Cho. In defense of lazy visual grounding for open-vocabulary semantic segmentation. In *European Conference on Computer Vision*, 2024. 2

[27] Markus Karmann and Onay Urfalioglu. Repurposing stable diffusion attention for training-free unsupervised interactive segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24518–24528, 2025. 1, 3

[28] Chanyoung Kim, Woojung Han, Dayun Ju, and Seong Jae Hwang. Eagle: Eigen aggregation learning for object-centric unsupervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 7, 8

[29] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 2

[30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 2

[31] Daiqing Li, Junlin Yang, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8300–8311, 2021. 2

[32] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3041–3050, 2023. 2

[33] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *Advances in Neural Information Processing Systems*, 2023. 2

[34] Marina Meila and Jianbo Shi. Learning segmentation by random walks. In *Advances in Neural Information Processing Systems*. MIT Press, 2000. 3, 4

[35] 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. 1, 2

[36] Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc-André Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Çağatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftekharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, José António Mariz, Raphael Meier, Sérgio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M. S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10): 1993–2024, 2015. 1, 4

[37] Koichi Namekata, Amirmojtaba Sabour, Sanja Fidler, and Seung Wook Kim. Emerdiff: Emerging pixel-level semantic knowledge in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 5, 7, 8, 1

[38] Mark Newman. *Networks*. Oxford University Press, 2018. Page 137, Eq: 6.31. 4

[39] Dantong Niu, Xudong Wang, Xinyang Han, Long Lian, Roei Herzig, and Trevor Darrell. Unsupervised universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22744–22754, 2024. 1, 7, 8

[40] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2023. 8

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, Los Alamitos, CA, USA, 2022. IEEE Computer Society. 1, 2

[42] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000. 1, 2, 3, 4, 5

[43] Leon Sick, Dominik Engel, Pedro Hermosilla, and Timo Ropinski. Unsupervised semantic segmentation through depth-guided feature correlation and sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3637–3646, 2024. 1, 5, 7, 8

[44] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 4

[45] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 1, 2, 3

[46] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse, attend, and segment: Unsupervised zero-shot segmentation using stable diffusion. *CVPR*, 2024. 1, 2, 3, 4, 5, 7, 8

[47] Nontawat Tritrong, Pitchaporn Rewatbowornwong, and Supasorn Suwajanakorn. Repurposing gans for one-shot semantic part segmentation. In *Proceedings of the IEEE/CVF*

conference on computer vision and pattern recognition, pages 4475–4485, 2021. 2

[48] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10748–10757, 2022. 2

[49] Jiabing Wang, Hong Peng, Jingsong Hu, and Chuangxin Yang. A graph clustering algorithm based on minimum and normalized cut. In *Computational Science–ICCS 2007: 7th International Conference, Beijing, China, May 27-30, 2007, Proceedings, Part I 7*, pages 497–504. Springer, 2007. 3, 5

[50] Xudong Wang, Rohit Girdhar, Stella X Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3124–3134, 2023. 2

[51] Yangtao Wang, Xi Shen, Yuan Yuan, Yuming Du, Maomao Li, Shell Xu Hu, James L Crowley, and Dominique Vaufreydaz. Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut. *IEEE transactions on pattern analysis and machine intelligence*, 2023. 1, 2

[52] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1206–1217, 2023. 2

[53] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[54] Xiao Zhang, David Yunis, and Michael Maire. Deciphering'what'and'where'visual pathways from spectral clustering of layer-distributed neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4165–4175, 2024. 3, 4

[55] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 5