FALCON: FAST PROXIMAL LINEARIZATION OF NORMALIZED CUTS FOR UNSUPERVISED IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Current zero-shot unsupervised segmentation methods based on normalized cuts (NCut) face three key limitations. First, they rely on recursive bipartitions with repeated eigen-decompositions, making them prohibitively expensive at scale. Second, each split requires spectral relaxation followed by rounding, introducing layers of approximation where the final partition may diverge from the true NCut objective. Third, existing heuristics lack convergence guarantees, and recursive bipartitioning offers no principled assurance of producing a stable K-way segmentation. We propose **Falcon**, a proximal-gradient solver that directly optimizes the discrete K-way NCut objective without spectral relaxation. We prove linear convergence in the number of tokens. Falcon computes closed-form gradient scores weighted by cluster volumes and performs row-wise one-hot proximal updates stabilized by inertia. A monotone backtracking scheme adaptively tunes the proximal parameter, ensuring non-decreasing NCut values. This design preserves discrete feasibility, removes repeated eigen-decomposition, and guarantees convergence under the Kurdyka–Łojasiewicz framework. Across six benchmarks, Falcon outperforms the strongest official baseline (DiffCut) by wide margins, e.g., +13.2 mIoU on VOC, +27.7 on COCO-Object, and +3.1 on Cityscapes, while remaining competitive on Pascal Context. It also runs up to an order of magnitude faster than recursive NCut. By pairing pretrained foundation models with a principled NCut solver, Falcon sets a new state of the art across six benchmarks and achieves the best performance on 17 of 18 benchmark–encoder pairs, underscoring both its robustness and its generality in bridging the gap between unsupervised and supervised segmentation.

1 INTRODUCTION

Semantic segmentation partitions an image into regions whose pixels share the same semantics (i.e. being part of the same object), which matters for AI systems in terms of perception, reasoning, planning, and acting in an object-centric manner Wang et al. (2023a); Zhu et al. (2016). As a critical and fundamental computer vision task, semantic segmentation underpins numerous downstream applications, including image editing, medical imaging, and autonomous driving Esser et al. (2023); Zhou et al. (2019b); Li et al. (2022b); Han et al. (2024). Semantic segmentation is a key computer vision task with countless downstream applications.

Semantic segmentation as rapidly improved over the last decade. Fully Convolutional Networks (FCNs) marked a major breakthrough in semantic segmentation Long et al. (2015). Subsequently, numerous improved architectures—such as SegNet Hu et al. (2018), U-Net Wang & Yang (2021), the DeepLab series Chen et al. (2017), and PSPNet Zhao et al. (2017)—were developed, with a focus on capturing multi-scale features and fusing low-level and high-level information for more accurate segmentation outcomes. The advent of instance segmentation and panoptic segmentation further broadened the applicability of segmentation tasks, with Mask R-CNN He et al. (2017) becoming a standard baseline Sapkota et al. (2024); Zhang et al. (2020). Following the success of the Vision Transformer (ViT) Dosovitskiy et al. (2020) in image classification, a variety of Transformer-based segmentation networks have emerged Oquab et al. (2023); Caron et al. (2021), offering enhanced global modeling capabilities and opening new research avenues in image segmentation, such as SegFormer Xie et al. (2021) Mask2Former Cheng et al. (2022). More recently, the Segment Anything Model (SAM) Ravi et al. (2024), which leverages large-scale data (1.1B segmentation annotations),

ViT-driven representations, and prompt-based segmentation, has introduced a new paradigm in computer vision. However, all these segmentation models require pixel-level annotations, which are both difficult and resource-intensive to obtain.

The difficulty to obtain good data has driven interest in unsupervised approaches, such as zero-shot unsupervised segmentation Tian et al. (2024); Couairon et al. (2025), where the goal is to segment images containing previously unseen categories—an inherently more challenging problem. Unsupervised image segmentation has recently advanced by incorporating self-supervised learning and traditional computer vision principles into deep learning pipelines Sick et al. (2024). Self-supervised learning (SSL) produces meaningful feature representations without requiring annotations Wen et al. (2022); Ziegler & Asano (2022). For instance, STEGO Hamilton et al. (2022) employs contrastive learning to extract patch-level features and refines segmentation masks through knowledge distillation and post-processing techniques such as Conditional Random Field (CRF). Another important approach incorporates traditional segmentation techniques such as clustering and graph-based optimization into deep learning pipelines Tian et al. (2024). Invariant Information Clustering (IIC) Ji et al. (2019) and PiCIE Cho et al. (2021) formulate segmentation as an unsupervised clustering problem, enforcing invariance and equivariance constraints to group pixels into semantically consistent regions. These algorithms use self supervised learning but no explicit cutting operations.

Other methods do use cutting operations. Graph-based segmentation has recently shown strong results when coupled with self-supervised features. TokenCut Wang et al. (2023b) and MaskCut Wang et al. (2023a) apply normalized cut (NCut) recursively on deep token embeddings, while DiffCut Couairon et al. (2025) augments graph optimization with diffusion priors to strengthen pixel connectivity and improve consistency. These pipelines typically adopt a spectral relaxation of NCut and perform hierarchical bipartitions—often guided by the Fiedler vector of the (normalized) Laplacian Shi & Malik (2000b); Ng et al. (2001). Despite their success, such recursive spectral schemes exhibit several limitations. *First*, the hierarchy is inherently greedy: once a split is made, earlier decisions cannot be globally revised, which may lead to suboptimal partitions. *Second*, the relax-and-round procedure introduces a relaxation gap; relying on a single eigenvector for binary decisions can trap the optimization in poor local configurations. *Third*, repeated eigen-decompositions across recursion levels incur nontrivial computational cost and slow inference.

To tackle these challenges, we propose Falcon, a fast proximal linearization of the NCut objective that optimizes the discrete assignment directly via row-wise one-hot projections under an exact gradient score, coupled with a monotone backtracking scheme. Instead of relying on recursive binary partitioning through the Fiedler vector, Falcon introduces a parallelizable K-way Normalized Cut formulation that effectively addresses the major limitations of existing graph-cut methods. Our contributions can be summarized as follows:

Method. We formulate NCut as a proximal-gradient solver that preserves discreteness throughout optimization. Each iteration computes closed-form row-wise one-hot updates from the exact gradient, stabilized by a proximal inertia term. This design eliminates repeated relaxations and eigen-decompositions, and aligns updates with the original fractional objective rather than heuristic surrogates. Our implementation is fully vectorized and memory-efficient.

Convergence. We prove that under the *Kurdyka–Łojasiewicz* (KL) property, the monotone backtracking scheme guarantees a non-decreasing objective, finite-length trajectories, and convergence to a critical point; if the KL exponent satisfies $\theta \leq \frac{1}{2}$, the local rate is at least linear.

Performance. Falcon sets new state of the art across six benchmarks and three pretrained encoders, achieving the best results on **17 of 18** benchmark–encoder pairs, underscoring its robustness and generality.

2 Related Works

Vision Foundation Models. Vision foundation models typically leverage unlabeled data to learn robust and generalizable representations. Early contrastive methods like MoCo He et al. (2020) and BYOL Grill et al. (2020) laid the groundwork for advanced frameworks such as SwAV Goyal et al. (2021), DINO Caron et al. (2021); Oquab et al. (2023); Darcet et al. (2023); Siméoni et al. (2025), and iBOT Zhou et al. (2021), while masked autoencoders He et al. (2022) have further refined reconstruction-based pre-training. Beyond purely visual approaches, multimodal pre-training has surged in prominence, with models like CLIP Radford et al. (2021), BLIP Li et al. (2022a), and SigLIP Zhai et al. (2023); Tschannen et al. (2025) aligning high-level image features to text. In parallel, diffusion-based methods such as Stable Diffusion Rombach et al. (2021); Gupta et al. (2024) extend these capabilities by learning rich generative representations, enabling tasks ranging

from zero-shot classification to semantic correspondence. Collectively, these developments highlight the efficacy of foundation models in scaling to large, diverse datasets and adapting readily to downstream tasks.

Semantic Segmentation. Semantic segmentation partitions an image into semantically coherent regions by labeling each pixel, enabling the understanding of the structured scene. It is broadly categorized into supervised and unsupervised methods. Supervised segmentation, extensively studied and achieving high accuracy Namekata et al. (2024); Wang et al. (2021); Cheng et al. (2021); Ravi et al. (2024), relies on large-scale annotated datasets. Recent work has explored text-based supervision to mitigate the need for dense annotations Ranasinghe et al. (2023); Xu et al. (2022); Cha et al. (2023); Ren et al. (2023). In contrast, unsupervised methods often require dataset-specific training to achieve competitive performance Liang et al. (2023); Feng et al. (2023); Cho et al. (2021); Li et al. (2023), and zero-shot segmentation for unseen categories remains challenging. DiffSeg Tian et al. (2024) leverages self-attention maps from a pre-trained diffusion model, applying KL-divergence-based iterative merging for segmentation. DiffCut Couairon et al. (2025) improves upon this by extracting richer encoder features from the self-attention block of a Transformer and incorporating a recursive N-Cut Shi & Malik (2000a) algorithm.

Graph-based Image Segmentation. Early approaches, such as Normalized Cut (N-Cut) Shi & Malik (2000a), optimized the segmentation problem through spectral graph theory, formalizing spectral clustering theory based on the N-Cut objective; yet, its computational limitations persisted. To improve efficiency and adaptability, F & P (2004) proposed an adaptive merging strategy that utilizes both intra-region and inter-region criteria, while Grady (2006) introduced a probabilistic random walk framework that leverages adjacent pixel relationships to handle complex textures and weak boundaries. Recent deep learning-integrated approaches, such as TokenCut Wang et al. (2022), AutoSC Fan et al. (2022), and DiffCut Couairon et al. (2025), compute token-level similarities via self-supervised Transformer features but remain constrained by N-Cut's recursive bisection strategy and hard segmentation constraints.

Proximal Gradient Methods. Proximal–gradient algorithms handle composite objectives that pair a smooth data term with a possibly nonsmooth regularizer frequently seen in learning (e.g., sparsity, total variation, simple constraints). A gradient step on the smooth part followed by a proximal step on the regularizer yields a simple, scalable routine with standard convergence guarantees and practical backtracking/line–search implementations Parikh & Boyd (2014). The accelerated variant FISTA achieves the optimal $\mathcal{O}(1/k^2)$ rate for convex problems, and monotone variants stabilize the objective along the iterates Beck & Teboulle (2009); Chambolle & Dossal (2015). For nonconvex formulations common in modern models, convergence to a critical point can be justified under the Kurdyka–Łojasiewicz framework and via blockwise/prox–linear updates Attouch et al. (2013); Bolte et al. (2014). Inertial extensions further improve empirical speed while retaining convergence guarantees under mild conditions Ochs et al. (2014).

3 METHODOLOGY

3.1 SEGMENTATION AS NCUT ON TOKENS

Graph-based segmentation casts image segmentation as partitioning a weighted graph. Following normalized-cut pipelines Wang et al. (2022; 2023a); Couairon et al. (2025), we build an undirected graph $G = (V, E)$ with N nodes, one per d-dimensional token from a vision Transformer:

$$V = \{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_N\}, \quad \boldsymbol{f}_i \in \mathbb{R}^d, \qquad \boldsymbol{F} = [\boldsymbol{f}_1 \quad \cdots \quad \boldsymbol{f}_N]^\top \in \mathbb{R}^{N \times d}. \tag{1}$$

We use a dense affinity (without self-loops) and form a nonnegative, symmetric similarity by row-normalizing features to unit ℓ_2 norm (cosine), rescaling to $[0, 1]$, and elementwise power sharpening:

$$\widehat{\boldsymbol{F}} = \mathrm{row\_norm}(\boldsymbol{F}), \quad \boldsymbol{S} = \widehat{\boldsymbol{F}}\widehat{\boldsymbol{F}}^\top, \quad \boldsymbol{W} = \phi(\boldsymbol{S})^{\odot\alpha},\ \alpha \geq 1, \qquad \mathrm{diag}(\boldsymbol{W}) \leftarrow \boldsymbol{0}, \tag{2}$$

where ϕ maps similarities to $[0, 1]$ (e.g., $\phi(s) = \max(s, 0)$ or min–max per image). We then define

$$\boldsymbol{d} = \boldsymbol{W}\boldsymbol{1}, \qquad \boldsymbol{D} = \mathrm{diag}(\boldsymbol{d}). \tag{3}$$

Let $\{P_1, \ldots, P_K\}$ be a partition and $\boldsymbol{x}_k \in \{0, 1\}^N$ its indicator with a row one-hot assignment $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K] \in \{0, 1\}^{N \times K}$ and $\boldsymbol{X}\boldsymbol{1} = \boldsymbol{1}$. The normalized cut objective Shi & Malik (2000b)

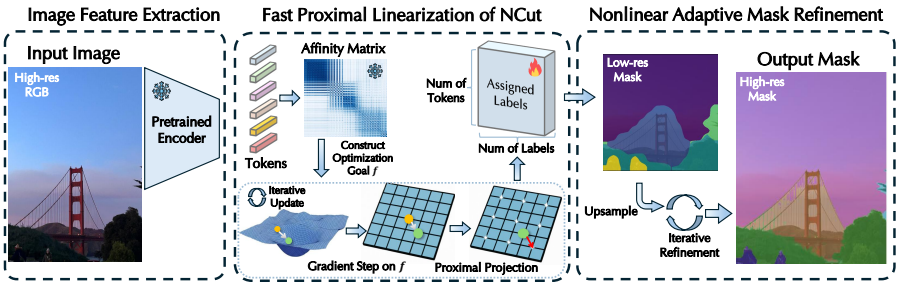

Figure 1: **Overview of Falcon.** (1) Image feature extraction: tokens are extracted from the input image. (2) Fast proximal linearization of NCut: an affinity matrix is constructed between tokens, and the proximal linearization process is iteratively updated to optimize semantic labels. (3) Nonlinear Adaptive Mask Refinement: the coarse mask is progressively refined, projecting the low-resolution labels onto the high-resolution mask using nonlinear pixel-affinity weights, as illustrated in Figure 2.

is

$$\mathrm{Ncut}(P_1,\dots,P_K)=\sum_{k=1}^{K}\frac{\mathrm{cut}(P_k,\bar{P}_k)}{\mathrm{vol}(P_k)},\quad \mathrm{cut}(P_k,\bar{P}_k)=\sum_{i\in P_k,\, j\notin P_k} W_{ij},\quad \mathrm{vol}(P_k)=\sum_{i\in P_k} d_i. \tag{4}$$

Equivalently, minimizing NCut is the same as maximizing the normalized association (i.e., a sum of K generalized Rayleigh quotients)

$$f(\boldsymbol{X})=\sum_{k=1}^{K}\frac{\boldsymbol{x}_k^\top \boldsymbol{W}\boldsymbol{x}_k}{\boldsymbol{x}_k^\top \boldsymbol{D}\boldsymbol{x}_k},\qquad \mathrm{Ncut}(\boldsymbol{X})=K-f(\boldsymbol{X}). \tag{5}$$

The problem is discrete and nonconvex; finding the exact optimum is NP-hard, which motivates principled relaxations and iterative schemes.

3.2 Falcon: Fast Proximal Linearization of NCut

Rather than relaxing one-hot rows, we model discreteness via an extended-valued composite:

$$\min_{\boldsymbol{X}\in\mathbb{R}^{N\times K}} \Phi(\boldsymbol{X})=h(\boldsymbol{X})+g(\boldsymbol{X}),\qquad h(\boldsymbol{X})=-f(\boldsymbol{X}),\qquad g(\boldsymbol{X})=\iota_{\mathcal{V}}(\boldsymbol{X}), \tag{6}$$

where $\mathcal{V}=\{\boldsymbol{X}\in\{0,1\}^{N\times K}:\boldsymbol{X}\mathbf{1}=\mathbf{1}\}$ is the row-wise one-hot simplex and $\iota_{\mathcal{V}}$ is its indicator (0 on $\mathcal{V}$, $+\infty$ otherwise). We additionally *enforce nonempty clusters algorithmically* (simple empty-cluster repair if they arise) and only require $v_k(\boldsymbol{X})>0$ *locally* in the analysis, where

$$\boldsymbol{d}=\boldsymbol{W}\mathbf{1},\qquad v_k(\boldsymbol{X})=\boldsymbol{x}_k^\top \boldsymbol{D}\boldsymbol{x}_k,\qquad \boldsymbol{D}=\mathrm{diag}(\boldsymbol{d}). \tag{7}$$

On $\{\boldsymbol{X}\in\mathcal{V}: v_k(\boldsymbol{X})>0,\forall k\}$, the map f is C^1, so $h=-f$ is smooth; unsmooth part g *exactly* encodes one-hot feasibility. No heuristic rounding is needed.

To conduct the gradient derivation of the smooth part, we introduce cached quantities

$$\boldsymbol{S}=\boldsymbol{W}\boldsymbol{X}\in\mathbb{R}^{N\times K},\qquad \boldsymbol{q}=(\boldsymbol{X}\odot\boldsymbol{S})^\top\mathbf{1}\in\mathbb{R}^{K},\qquad \boldsymbol{v}=\boldsymbol{X}^\top\boldsymbol{d}\in\mathbb{R}^{K}, \tag{8}$$

so that $q_k=\boldsymbol{x}_k^\top\boldsymbol{W}\boldsymbol{x}_k$ and $v_k=\boldsymbol{x}_k^\top\boldsymbol{D}\boldsymbol{x}_k$. By the quotient rule,

$$\nabla_{\boldsymbol{x}_k}\Big(\tfrac{q_k}{v_k}\Big)=\tfrac{2}{v_k^2}\big(v_k\,\boldsymbol{W}\boldsymbol{x}_k-q_k\,\boldsymbol{D}\boldsymbol{x}_k\big),\quad\Rightarrow\quad \nabla f(\boldsymbol{X})=2\big(\boldsymbol{S}\,\boldsymbol{v}^{-T}-(\boldsymbol{D}\boldsymbol{X})\,\boldsymbol{c}^\top\big),\ \ c_k=\tfrac{q_k}{v_k^2}. \tag{9}$$

Here $\boldsymbol{v}^{-T}$ denotes column-wise division by $\boldsymbol{v}$.

For the unsmooth part g, Falcon applies proximal projection onto discrete assignments. Let $\boldsymbol{X}^{(t)}$ be the current iterate. If ∇f is locally L_t-Lipschitz in a neighborhood where $v_k>0$, then for any $\tau_t\ge L_t$ the standard smoothness inequality gives a local quadratic *minorizer* of f:

$$f(\boldsymbol{Y})\ \ge\ f(\boldsymbol{X}^{(t)})+\langle\nabla f(\boldsymbol{X}^{(t)}),\boldsymbol{Y}-\boldsymbol{X}^{(t)}\rangle-\tfrac{\tau_t}{2}\|\boldsymbol{Y}-\boldsymbol{X}^{(t)}\|_F^2. \tag{10}$$

Maximizing this surrogate *over the original discrete set* $\mathcal{V}$ is row-separable. Because each feasible row is one-hot, $\|\boldsymbol{y}_i\|_2^2 \equiv 1$ is constant and $-\frac{\tau_t}{2}\|\boldsymbol{y}_i - \boldsymbol{X}_i^{(t)}\|_2^2 = \frac{\tau_t}{2}(2\langle \boldsymbol{y}_i, \boldsymbol{X}_i^{(t)}\rangle - 1)$, so only the cross term remains. This yields the closed-form argmax update

$$\boldsymbol{x}_i^{(t+1)} \;=\; e_{\,\arg\max_{k\in[K]}\; \mu_{ik}^{(t)} + \tau_t X_{ik}^{(t)}}, \qquad \boldsymbol{\mu}^{(t)} = \nabla f(\boldsymbol{X}^{(t)}), \tag{11}$$

where ties are broken deterministically (e.g. smallest k). The additive $\tau_t X_{ik}^{(t)}$ acts as inertia, preventing gratuitous flips while preserving feasibility.

Because L_t is unknown and may vary, we adopt a monotone line search on τ_t: start with $\tau_t = \tau_0 > 0$; compute equation 11; *accept* if

$$f(\boldsymbol{X}^{(t+1)}) \;\geq\; f(\boldsymbol{X}^{(t)}) \;+\; \frac{\delta\,\tau_t}{2}\,\|\boldsymbol{X}^{(t+1)} - \boldsymbol{X}^{(t)}\|_F^2, \qquad \delta \in (0,1), \tag{12}$$

otherwise set $\tau_t \leftarrow \gamma\tau_t$ ($\gamma > 1$) and recompute. This Armijo-type sufficient *increase* guarantees monotonic ascent of f and finite termination of the backtracking. Equivalently, in the minimization view for $\Phi(\boldsymbol{X}) = -f(\boldsymbol{X}) + \iota_{\mathcal{V}}(\boldsymbol{X})$, equation 12 is identical to

$$\Phi(\boldsymbol{X}^{(t+1)}) \;\leq\; \Phi(\boldsymbol{X}^{(t)}) \;-\; \frac{\delta\,\tau_t}{2}\,\|\boldsymbol{X}^{(t+1)} - \boldsymbol{X}^{(t)}\|_F^2,$$

which we use in the Kurdyka–Łojasiewicz based analysis (Appendix C).

Convergence under the Kurdyka–Łojasiewicz (KL) Property. Falcon is a forward–backward scheme with monotone backtracking: the update equation 11 is the proximal step for $g = \iota_{\mathcal{V}}$ (row-wise projection), driven by the linearization of $h = -f$, while equation 12 provides a quantitative acceptance rule. Under standard conditions—bounded sublevel sets, local Lipschitz continuity of ∇f in a positive-volume neighborhood, and the KL property of Φ at a limit point—the generated sequence enjoys: (i) finite termination of backtracking and sufficient descent of Φ; (ii) a uniform subgradient–step bound; (iii) (iii) global convergence of the whole sequence[1]; and (iv) linear rates when the KL exponent satisfies $\theta \leq \frac{1}{2}$. The precise assumptions, lemmas, and proof (following Jia et al. (2023), Section 4) are given in Appendix C, with the notational correspondence $\tau_t \;\leftrightarrow\; \gamma_k$ and the same δ. Our discrete NCut objective involves polynomial expressions of the affinity matrix together with one-hot constraints on cluster assignments, which makes it a semi-algebraic function. Semi-algebraic functions are known to satisfy the KL property with an exponent $\theta \leq \frac{1}{2}$ Bolte et al. (2007); Attouch et al. (2013). Consequently, our proximal-gradient scheme enjoys at least a local linear convergence rate near stationary points, in addition to the monotonic descent guaranteed by the backtracking line search.

3.3 Segmentation Mask Generation

The optimization in Section 3.2 yields a *discrete* token-level assignment $\ell \in \{1,\dots,K\}^{h\times w}$ on the token grid. While this already provides a coarse mask, its resolution (h, w) is significantly lower than the image. We therefore adopt a two-stage refinement to an *intermediate* grid of size (H, W) (e.g., 128×128), which densifies the partition while remaining computationally light.

We upsample the discrete label map by nearest-neighbor interpolation to obtain $\hat{\ell} \in \{1,\dots,K\}^{H\times W}$. In parallel, the feature map is bilinearly upsampled to $Z \in \mathbb{R}^{C\times H\times W}$; we denote by $\mathbf{z}_{h,w} \in \mathbb{R}^C$ the feature embedding at location (h, w) on the intermediate grid. The nearest-neighbor step preserves sharp boundaries in the label field, whereas bilinear interpolation maintains smooth variation in the embedding space.

From $\hat{\ell}$ we form the one-hot mask tensor $M \in \{0,1\}^{K\times H\times W}$ with $M_{k,h,w} = 1$ iff $\hat{\ell}_{h,w} = k$. The prototype of partition k is the average embedding of its assigned locations:

$$\mathbf{c}_k \;=\; \frac{\sum_{h=1}^{H}\sum_{w=1}^{W} M_{k,h,w}\,\mathbf{z}_{h,w}}{\sum_{h=1}^{H}\sum_{w=1}^{W} M_{k,h,w}}, \qquad \mathbf{c}_k \in \mathbb{R}^C. \tag{13}$$

[1]Here, "global convergence" refers to the convergence of the entire sequence to a stationary point, rather than convergence to a global optimum of the NP-hard NCut objective.

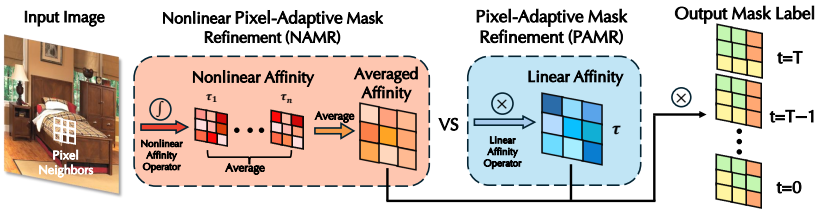

Figure 2: **NAMR vs PAMR.** The Nonlinear Pixel-Adaptive Mask Refinement (NAMR) computes pixel affinities using a nonlinear operator and averages them across multiple temperatures. In contrast, Pixel-Adaptive Mask Refinement (PAMR) uses a linear operator and computes affinities at a single temperature. After affinities are obtained, mask labels are iteratively updated based on them.

These prototypes summarize high-resolution cluster statistics on the (H, W) grid. Each location is then reassigned to the prototype with the largest dot-product similarity:

$$\ell^*_{h,w} \;=\; \arg\max_{k\in[K]} \; \mathbf{z}_{h,w}^{\top}\mathbf{c}_k, \tag{14}$$

yielding the refined mask $\ell^* \in \{1,\dots,K\}^{H\times W}$. This intermediate-resolution refinement reduces tokenization artifacts and enforces spatial coherence, and can be trivially upsampled to the original image size if desired.

3.4 Mask Refinement from Token Maps to Pixel-Level Masks

Our segmentation masks are first computed at the transformer-token grid by a graph-cut module; consequently, the output resolution is limited by the token stride. To recover pixel-level masks aligned with image boundaries, we apply a light-weight refinement step on top of a nearest-neighbor (NN) upsampled initialization.

Classical refiners, such as dense CRFs or matting-based solvers, can improve boundary adherence, but they are comparatively expensive, often CPU-bound, or require solving global systems, which complicates end-to-end training due to backpropagation and memory costs.

Pixel-Adaptive Mask Refinement (PAMR). Let $I \in \mathbb{R}^{H\times W\times 3}$ be the RGB image and $M_0 \in [0,1]^{(C+1)\times H\times W}$ the NN-upsampled class-probability tensor (including background). For a pixel (i,j) and a local neighborhood $\mathcal{N}(i,j)$ (built from small 3×3 kernels with dilations), PAMR forms per-direction affinities by a softmax over negative, locally normalized RGB differences:

$$\alpha_{i,j,l,n} = \frac{\exp\big(-\bar{r}(I_{i,j}, I_{l,n})/(\varepsilon+\sigma_{i,j})\big)}{\sum_{(q,r)\in\mathcal{N}(i,j)}\exp\big(-\bar{r}(I_{i,j}, I_{q,r})/(\varepsilon+\sigma_{i,j})\big)}, \quad \bar{r}(I_{i,j}, I_{l,n}) = \tfrac{1}{3}\sum_{k=1}^{3}\big|I_k(i,j) - I_k(l,n)\big|,$$

where $\sigma_{i,j}$ is the local standard deviation of $\bar{r}$ around (i,j) and $\varepsilon > 0$ ensures numerical stability. Refinement proceeds by T iterations of locally weighted averaging (one-step message passing):

$$M^t_{:,i,j} = \sum_{(l,n)\in\mathcal{N}(i,j)} \alpha_{i,j,l,n}\, M^{t-1}_{:,l,n}, \qquad t = 1,\dots,T, \quad M^0 := M_0.$$

Each update is a convex combination across neighbors (row-stochastic weights), yielding an anisotropic diffusion that respects intensity edges and is stable under small T.

Non-linear Adaptive Mask Refinement (NAMR). To demonstrate that our Falcon pipeline is agnostic to the refiner choice, we additionally instantiate a non-linear variant. NAMR replaces $\bar{r}$ with a contrast-amplified discrepancy $\bar{r}^{\mathrm{nl}}$ obtained by a pointwise nonlinearity ϕ (e.g., $\phi(x) = x + 1.5\,\mathrm{ELU}(x)$) before taking absolute values and channel averaging, and aggregates multiple "temperatures" $\tau \in \mathcal{T}$:

$$\alpha^{(\tau)}_{i,j,l,n} = \mathrm{softmax}_{(l,n)\in\mathcal{N}(i,j)}\left(-\frac{\bar{r}^{\mathrm{nl}}(I_{i,j}, I_{l,n})}{\varepsilon + \tau\,\sigma_{i,j}}\right), \quad M_T^{(\tau)} = \underbrace{\mathsf{MP}^{T}_{\alpha^{(\tau)}}(M_0)}_{T \text{ message-passing steps}}, \bar{M}_T = \tfrac{1}{|\mathcal{T}|}\sum_{\tau\in\mathcal{T}} M_T^{(\tau)}.$$

Small τ emphasizes edge preservation; larger τ favors smoothing; their average improves robustness across textures and scales. In experiments, we compare three choices: *NN only*, *PAMR*, and *NAMR*.

Table 1: **Unsupervised segmentation on six benchmarks (higher is better).** Reports **mIoU**. MaskCLIP Dong et al. (2023) requires text prompts. †: reported by us. AutoSC, DiffCut, and Falcon use SSD-1B encoder.

Method	mIoU (%) ↑					
	VOC	Context	COCO-Object	COCO-Stuff-27	Cityscapes	ADE20K
MaskCLIP Dong et al. (2023)	38.80	23.60	20.60	19.60	10.00	9.80
MaskCut Wang et al. (2023a)	53.80	43.40	30.10	41.70	18.70	35.70
DiffSeg Tian et al. (2024)	49.80	48.80	23.20	44.20	16.80	37.70
DiffCut Couairon et al. (2025)	65.20	56.50	34.10	49.10	30.60	44.30
AutoSC† (ours, refined)	77.57	57.27	61.56	49.39	25.72	40.10
DiffCut† (ours, refined)	71.68	**58.17**	61.65	49.18	30.77	44.40
Falcon (ours)	**78.40**	57.15	**61.80**	**50.37**	**33.69**	**45.17**

4 EXPERIMENTS

4.1 IMPLEMENTATION DETAILS

Datasets and Metrics. We evaluate Falcon on six widely used benchmarks spanning objects, scenes, and urban environments: Pascal VOC Everingham et al. (2015) (20 classes), Pascal Context Mottaghi et al. (2014) (59 classes with contextual labels), COCO-Object Lin et al. (2014) (80 categories), COCO-Stuff-27 Lin et al. (2014) (27 consolidated classes), Cityscapes Cordts et al. (2016) (27 urban categories), and ADE20K Zhou et al. (2019a) (150 classes). Segmentation quality is measured by mean intersection-over-union (mIoU). Since Falcon produces class-agnostic masks, we apply the Hungarian algorithm Kuhn (1955) to align predictions with ground-truth labels, using a many-to-one mapping for background categories. As a secondary metric, we report Pixel Accuracy (Pixel Acc.), the proportion of correctly classified pixels after the same alignment procedure.

Settings. All images are resized to 1024×1024. Features are extracted at lower resolutions: 32×32 from diffusion models Gupta et al. (2024) (prompt-free, t=50) and 64×64 from DINOv3 Siméoni et al. (2025), followed by ℓ_2 normalization and α power transformation. Falcon then produces an initial segmentation, which is refined by upsampling the cluster assignments to 128×128 and then recomputing labels with the corresponding features via specific refinements. The number of segments K is determined via spectral thresholding: we count eigenmodes below a negative threshold κ and clamp K to a predefined range. Experiments are run in PyTorch on a single NVIDIA RTX 4090. This unified pipeline supports all evaluated benchmarks.

Mask refinement. We adopt standard settings for **PAMR** (as in DiffCut): 3×3 kernels with dilations $D = \{1, 2, 4, 8\}$, yielding $P = 8|D|$ directional filters, $T = 10$ refinement steps, per-pixel local variance $\sigma_{i,j}$ over $\mathcal{N}(i, j)$, and $\varepsilon = 10^{-6}$. Gradients are not propagated through the refiner. For **NAMR**, we use the same neighborhoods and $T = 10$, define $\phi(x) = x + 1.5\,\mathrm{ELU}(x)$ to construct $\bar{r}^{\mathrm{nl}}$, and average over a temperature set $\mathcal{T} = \{0.06, 0.08, 0.10, 0.12, 0.14, 0.16, 0.18\}$ (seven scales). Unless noted, no additional data-fidelity mixing with M_0 is applied. The **NN** baseline simply upsamples token masks to (H, W) via nearest-neighbor interpolation without refinement.

4.2 MAIN RESULTS

Overall performance. Table 1 reports mIoU on six standard benchmarks. Unless noted otherwise, AutoSC, DiffCut, and Falcon numbers use the SSD-1B encoder with PAMR refinement (matching the settings in Table 2).

Relative to the strongest official baseline, DiffCut Couairon et al. (2025), Falcon establishes a new state of the art on five out of six datasets: VOC (+13.2, 78.40 vs. 65.2), COCO-Object (+27.7, 61.80 vs. 34.1), COCO-Stuff-27 (+1.3, 50.37 vs. 49.1), Cityscapes (+3.1, 33.69 vs. 30.6), and ADE20K (+0.9, 45.17 vs. 44.3); on Pascal Context, Falcon remains competitive (57.15 vs. 56.5). These gains confirm that Falcon improves substantially over the previous state-of-the-art.

We further include refined re-implementations of DiffCut and AutoSC (†) for a stronger comparison. While these enhanced baselines outperform their official counterparts (e.g., DiffCut† reaches 61.65 on COCO-Object and 58.17 on Context), Falcon still surpasses them in most cases: +0.83 on VOC, +0.98 on COCO-Stuff-27, +2.92 on Cityscapes, and +0.77 on ADE20K. Averaged over all six

Table 2: **Unsupervised segmentation across encoders (higher is better).** Columns report **Pixel Acc. (%)** and **mIoU (%)** for each encoder.

Settings			SSD-1B		DINOv3-B		SD2.1	
Benchmark	**Method**	**Refin.**	**Pixel Acc. (%)**	**mIoU (%)**	**Pixel Acc. (%)**	**mIoU (%)**	**Pixel Acc. (%)**	**mIoU (%)**
Cityscapes	DiffCut† (ours, refined)	none	83.55	28.35	28.38	12.75	57.29	17.09
		pamr	85.91	30.77	29.94	12.02	67.26	19.41
		namr	86.02	30.95	29.48	12.37	65.40	19.41
	Falcon	none	83.08	30.56	63.26	**25.29**	78.56	26.65
		pamr	83.74	**33.69**	63.85	24.40	82.81	28.04
		namr	83.65	33.50	63.88	24.41	82.87	**28.21**
VOC	DiffCut† (ours, refined)	none	81.60	68.40	76.71	59.41	70.41	50.88
		pamr	81.81	71.68	76.66	65.73	71.57	60.41
		namr	82.04	71.94	76.54	65.63	71.40	60.16
	Falcon	none	88.12	**79.15**	84.48	**76.44**	86.78	**77.94**
		pamr	87.97	78.40	83.79	75.27	86.51	77.10
		namr	88.28	78.83	84.19	75.62	86.78	77.35
Context	DiffCut† (ours, refined)	none	77.86	54.55	61.14	37.69	70.45	45.66
		pamr	80.36	**58.17**	61.16	42.63	74.02	52.05
		namr	80.25	58.10	61.24	43.04	73.57	52.02
	Falcon	none	78.47	54.90	62.08	44.24	75.53	52.38
		pamr	80.53	57.15	62.22	45.00	78.04	**55.56**
		namr	80.48	57.23	62.31	**45.21**	77.84	55.39
COCO-Stuff	DiffCut† (ours, refined)	none	74.97	46.21	23.35	13.07	60.73	32.41
		pamr	78.47	49.18	24.43	13.54	67.37	37.43
		namr	78.19	49.05	24.08	13.40	66.18	36.49
	Falcon	none	75.85	47.88	48.39	**30.43**	72.17	43.64
		pamr	78.56	**50.37**	48.82	30.32	75.07	**46.41**
		namr	78.39	50.28	48.80	**30.43**	74.64	46.18
COCO-Object	DiffCut† (ours, refined)	none	80.28	62.96	72.42	52.71	67.36	48.82
		pamr	77.44	61.65	71.74	59.11	67.93	54.70
		namr	78.60	62.94	71.75	60.55	67.88	55.53
	Falcon	none	84.67	**63.98**	69.55	**62.19**	76.17	59.21
		pamr	81.22	61.80	67.70	59.69	73.76	58.21
		namr	82.30	62.74	68.31	60.78	74.68	**59.41**
ADE20K	DiffCut† (ours, refined)	none	69.55	42.54	28.05	26.50	60.43	38.38
		pamr	72.15	44.40	29.23	25.85	65.82	41.92
		namr	71.84	44.53	28.86	26.32	64.71	41.56
	Falcon	none	67.64	43.05	59.15	41.87	68.33	41.11
		pamr	70.85	**45.17**	60.20	42.29	70.73	42.62
		namr	70.38	45.06	60.04	**42.44**	70.42	**42.70**

datasets, Falcon attains **54.43** mIoU, exceeding the strongest per-dataset baseline by **+0.77** points on average.

Compared to earlier prompt-free approaches (MaskCut, DiffSeg) and the optimized DiffCut/AutoSC re-implementations, Falcon consistently delivers robust improvements across both object-centric and scene-centric settings.

Across encoders and refinement. Table 2 analyzes robustness across feature backbones (SSD-1B, DINOv3-B, SD2.1) and refinement choices (None/PAMR/NAMR). SSD- 1B yields the strongest overall results and serves as our default. Refinement effects are dataset-dependent: on scene-heavy datasets (Cityscapes, COCO-Stuff-27, ADE20K), PAMR/NAMR provide consistent gains (e.g., Cityscapes 33.69 with PAMR; COCO-Stuff-27 46.41 with PAMR on SD2.1), whereas on object-centric VOC, the plain segmentation is already very strong (79.15 without refinement). DINOv3-B and SD2.1 trail SSD- 1B but preserve the same ordering across refinement options, indicating encoder-agnostic behavior of Falcon.

Falcon delivers state-of-the-art unsupervised segmentation on five of six benchmarks with a single, prompt-free pipeline, and remains competitive on the remaining dataset. Its gains persist across diverse encoders and refinement schemes, underscoring both effectiveness and robustness.

4.3 Runtime and Efficiency

Table 3 and the stacked-bar plots report the end-to-end evaluation time across *Data Preparation*, *Feature Extraction*, and *Mask Generation*. Since the first two stages are identical for all methods, runtime differences stem almost entirely from the segmentation solver.

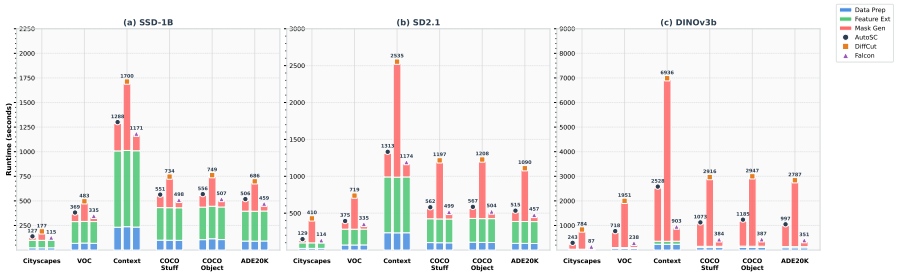

Figure 3: **The comparison of total evaluation time on various datasets.** Falcon can largely shorten inference time than recursive N-Cut on a single RTX4090.

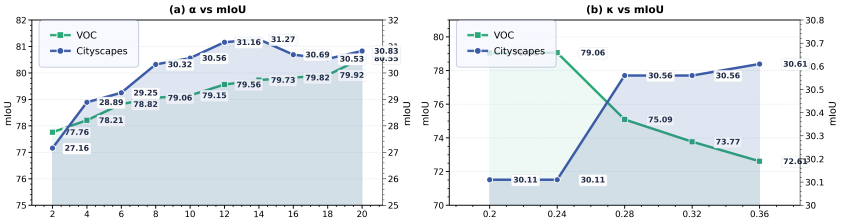

Figure 4: **Hyperparameter sensitivity.** Impact of the power parameter α (left) and spectral cutoff κ (right) on mIoU for Cityscapes and VOC.

Falcon consistently delivers the fastest Mask Generation time, yielding the lowest overall Total Time. The speedup is particularly striking when token counts are high: for example, on *DINOv3-B, Cityscapes*, Falcon reduces Total Time from 784.04s (DiffCut) to 87.47s, and Mask Generation from 747.97s to 52.49s. These results demonstrate that Falcon scales gracefully with token resolution, avoiding the explosive runtime growth of recursive methods. The observed convergence speed is fully consistent with our KL-based theoretical analysis, confirming both the efficiency and robustness of the approach.

4.4 Hyperparameter Tuning Studies

Power Transformation in the Affinity Matrix. In high-dimensional embedding spaces, graph-based methods often suffer from similarity collapse, where pairwise affinities become nearly uniform, thereby blurring true cluster boundaries. To address this, we apply a power transformation that nonlinearly rescales similarities: stronger affinities are amplified while weaker ones are suppressed. This contrast enhancement improves the spectral embedding, reduces sensitivity to noise, and yields more stable partitions. As shown in Figure 4, tuning the power parameter α leads to promised gains in mean Intersection-over-Union (mIoU), validating the effectiveness of this transformation in clarifying semantic structure.

Spectral Threshold for K Selection. The number of segments K is estimated via spectral thresholding: eigenmodes below a negative cutoff κ are counted, and the resulting K is clamped to a predefined range. Varying κ does affect segmentation quality, but the effect is dataset-dependent and generally moderate. As shown in Figure 4, Cityscapes and Pascal VOC exhibit slightly different sensitivities to κ, yet overall performance remains relatively stable across a reasonable range of values.

5 Conclusion

We introduced Falcon, a proximal-gradient solver for the discrete NCut objective that avoids spectral relaxation, preserves feasibility, and enjoys convergence guarantees under the *Kurdyka–Łojasiewicz* framework. Across six benchmarks and three pretrained encoders, Falcon achieves state-of-the-art results on 17 of 18 benchmark–encoder pairs, while running up to an order of magnitude faster than recursive NCut methods. These results demonstrate the benefits of treating NCut as a principled optimization problem and point toward broader opportunities for combining foundation models with discrete solvers in segmentation and beyond.

REFERENCES

Hedy Attouch, Jérôme Bolte, and Benar F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1–2):91–129, 2013. doi: 10.1007/s10107-011-0484-9.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. doi: 10.1137/080716542.

Jérôme Bolte, Aris Daniilidis, Adrian S. Lewis, and Masahiro Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572, 2007. doi: 10.1137/060670298.

Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1–2):459–494, 2014. doi: 10.1007/s10107-013-0701-9.

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*, pp. 9650–9660, 2021.

Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11165–11174, 2023.

Antonin Chambolle and Charles Dossal. On the convergence of the iterates of the FISTA algorithm. *Journal of Optimization Theory and Applications*, 166(3):968–982, 2015. doi: 10.1007/s10957-015-0746-4.

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.

Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1290–1299, 2022.

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*, pp. 16794–16804, 2021.

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*, pp. 3213–3223, 2016.

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, 2025.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2023.

Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10995–11005, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7346–7356, 2023.

Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015.

Felzenszwalb P F and Huttenlocher D P. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 2004.

Jicong Fan, Yiheng Tu, Zhao Zhang, Mingbo Zhao, and Haijun Zhang. A simple approach to automated spectral clustering. NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Qianli Feng, Raghudeep Gadde, Wentong Liao, Eduard Ramon, and Aleix Martinez. Network-free, unsupervised semantic segmentation with synthetic images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23602–23610, 2023.

Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, et al. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988*, 2021.

L. Grady. Random walks for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Yatharth Gupta, Vishnu V Jaddipal, Harish Prabhala, Sayak Paul, and Patrick Von Platen. Progressive knowledge distillation of stable diffusion xl using layer level loss. *arXiv preprint arXiv:2401.02677*, 2024.

Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T Freeman. Unsupervised semantic segmentation by distilling feature correspondences. *arXiv preprint arXiv:2203.08414*, 2022.

Xiangyu Han, Zhen Jia, Boyi Li, Yan Wang, Boris Ivanovic, Yurong You, Lingjie Liu, Yue Wang, Marco Pavone, Chen Feng, et al. Extrapolated urban view synthesis benchmark. *arXiv preprint arXiv:2412.05256*, 2024.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

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*, pp. 9865–9874, 2019.

Xiaoxi Jia, Christian Kanzow, and Patrick Mehlitz. Convergence analysis of the proximal gradient method in the presence of the kurdyka–Łojasiewicz property without global lipschitz assumptions. *SIAM Journal on Optimization*, 33(4):3038–3056, 2023. doi: 10.1137/23M1548293. URL https://doi.org/10.1137/23M1548293.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022a.

Kehan Li, Zhennan Wang, Zesen Cheng, Runyi Yu, Yian Zhao, Guoli Song, Chang Liu, Li Yuan, and Jie Chen. Acseg: Adaptive conceptualization for unsupervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7162–7172, 2023.

Qi Li, Yue Wang, Yilun Wang, and Hang Zhao. Hdmapnet: An online hd map construction and evaluation framework. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 4628–4634. IEEE, 2022b.

Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7061–7070, 2023.

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 *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pp. 740–755. Springer, 2014.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 891–898, 2014.

Koichi Namekata, Amirmojtaba Sabour, Sanja Fidler, and Seung Wook Kim. Emerdiff: Emerging pixel-level semantic knowledge in diffusion models. *arXiv preprint arXiv:2401.11739*, 2024.

Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001.

Peter Ochs, Yunjin Chen, Thomas Brox, and Thomas Pock. ipiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2):1388–1419, 2014. doi: 10.1137/130942954.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1 (3):127–239, 2014. doi: 10.1561/2400000003.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, and Jonathon Shlens. Perceptual grouping in contrastive vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5571–5584, 2023.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

Pengzhen Ren, Changlin Li, Hang Xu, Yi Zhu, Guangrun Wang, Jianzhuang Liu, Xiaojun Chang, and Xiaodan Liang. Viewco: Discovering text-supervised segmentation masks via multi-view semantic consistency. *arXiv preprint arXiv:2302.10307*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

Ranjan Sapkota, Dawood Ahmed, and Manoj Karkee. Comparing yolov8 and mask r-cnn for instance segmentation in complex orchard environments. *Artificial Intelligence in Agriculture*, 13: 84–99, 2024.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000a.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000b.

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*, pp. 3637–3646, 2024.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3554–3563, 2024.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5463–5474, 2021.

S Wang and F Yang. Remote sensing image semantic segmentation method based on u-net feature fusion optimization strategy. *Comput. Sci*, 48(8):162–168, 2021.

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*, pp. 3124–3134, 2023a.

Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L. Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14543–14553, June 2022.

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*, 45(12):15790–15801, 2023b.

Xin Wen, Bingchen Zhao, Anlin Zheng, Xiangyu Zhang, and Xiaojuan Qi. Self-supervised visual representation learning with semantic grouping. *Advances in neural information processing systems*, 35:16423–16438, 2022.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18134–18144, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Yiqing Zhang, Jun Chu, Lu Leng, and Jun Miao. Mask-refined r-cnn: A network for refining object details in instance segmentation. *Sensors (Basel, Switzerland)*, 20, 2020. URL `https://api.semanticscholar.org/CorpusID:211191810`.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890, 2017.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019a.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, 39(6):1856–1867, 2019b.

Hongyuan Zhu, Fanman Meng, Jianfei Cai, and Shijian Lu. Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation. *Journal of Visual Communication and Image Representation*, 34:12–27, 2016.

Adrian Ziegler and Yuki M Asano. Self-supervised learning of object parts for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14502–14511, 2022.

A Equivalence of K-way NCut and a Sum of Generalized Rayleigh Quotients

We consider an undirected weighted graph with a symmetric nonnegative affinity matrix $\boldsymbol{W} \in \mathbb{R}^{N\times N}$ and degree matrix $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{W}\mathbf{1})$, and let $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$ be the combinatorial Laplacian. A K-way partition $\{P_1, \dots, P_K\}$ is represented by one-hot indicator columns $\boldsymbol{x}_k \in \{0,1\}^N$ and the assignment matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_K] \in \{0,1\}^{N\times K}$ satisfying $\boldsymbol{X}\mathbf{1} = \mathbf{1}$.

For each part P_k, denote its volume and cut by

$$\mathrm{vol}(P_k) = \sum_{i\in P_k} d_i, \qquad \mathrm{cut}(P_k, \bar{P}_k) = \sum_{i\in P_k,\, j\notin P_k} W_{ij}. \tag{15}$$

The normalized cut objective Shi & Malik (2000b) is

$$\mathrm{Ncut}(P_1, \dots, P_K) = \sum_{k=1}^{K} \frac{\mathrm{cut}(P_k, \bar{P}_k)}{\mathrm{vol}(P_k)}. \tag{16}$$

With the indicator $\boldsymbol{x}_k$ of P_k, we have the following standard equalities:

$$\mathrm{vol}(P_k) = \boldsymbol{x}_k^\top \boldsymbol{D}\boldsymbol{x}_k, \qquad \mathrm{cut}(P_k, \bar{P}_k) = \boldsymbol{x}_k^\top \boldsymbol{L}\boldsymbol{x}_k = \boldsymbol{x}_k^\top \boldsymbol{D}\boldsymbol{x}_k - \boldsymbol{x}_k^\top \boldsymbol{W}\boldsymbol{x}_k. \tag{17}$$

Since $[\boldsymbol{x}_k]_i = 1$ iff $i \in P_k$, $\boldsymbol{x}_k^\top \boldsymbol{D}\boldsymbol{x}_k = \sum_{i=1}^N d_i [\boldsymbol{x}_k]_i^2 = \sum_{i\in P_k} d_i = \mathrm{vol}(P_k)$. Moreover, using the well-known Laplacian quadratic form $\boldsymbol{z}^\top \boldsymbol{L}\boldsymbol{z} = \frac{1}{2}\sum_{i,j} W_{ij}(z_i - z_j)^2$ and taking $\boldsymbol{z} = \boldsymbol{x}_k \in \{0,1\}^N$, the summand is 1 iff $\{i, j\}$ crosses the cut $(P_k, \bar{P}_k)$, which yields $\boldsymbol{x}_k^\top \boldsymbol{L}\boldsymbol{x}_k = \mathrm{cut}(P_k, \bar{P}_k)$. The decomposition $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$ gives the alternative expression. □

Dividing equation 17 termwise gives

$$\frac{\mathrm{cut}(P_k, \bar{P}_k)}{\mathrm{vol}(P_k)} = \frac{\boldsymbol{x}_k^\top \boldsymbol{L}\boldsymbol{x}_k}{\boldsymbol{x}_k^\top \boldsymbol{D}\boldsymbol{x}_k} = 1 - \frac{\boldsymbol{x}_k^\top \boldsymbol{W}\boldsymbol{x}_k}{\boldsymbol{x}_k^\top \boldsymbol{D}\boldsymbol{x}_k}. \tag{18}$$

Summing equation 18 over $k = 1, \ldots, K$ and using equation 16 yields

$$\mathrm{Ncut}(\boldsymbol{X}) = \sum_{k=1}^{K} \frac{\boldsymbol{x}_k^\top \boldsymbol{L}\boldsymbol{x}_k}{\boldsymbol{x}_k^\top \boldsymbol{D}\boldsymbol{x}_k} = K - \sum_{k=1}^{K} \frac{\boldsymbol{x}_k^\top \boldsymbol{W}\boldsymbol{x}_k}{\boldsymbol{x}_k^\top \boldsymbol{D}\boldsymbol{x}_k}. \tag{19}$$

Define the normalized association

$$f(\boldsymbol{X}) \;=\; \sum_{k=1}^{K} \frac{\boldsymbol{x}_k^\top \boldsymbol{W}\boldsymbol{x}_k}{\boldsymbol{x}_k^\top \boldsymbol{D}\boldsymbol{x}_k}, \qquad \text{so that} \qquad \mathrm{Ncut}(\boldsymbol{X}) \;=\; K - f(\boldsymbol{X}). \tag{20}$$

Each summand of $f(\boldsymbol{X})$ is a generalized Rayleigh quotient of the matrix pencil $(\boldsymbol{W}, \boldsymbol{D})$ evaluated at the indicator $\boldsymbol{x}_k$. With K fixed, minimizing Ncut is equivalent to maximizing f over the discrete assignment set $\mathcal{X} := \{\boldsymbol{X} \in \{0,1\}^{N\times K} : \boldsymbol{X}\mathbf{1} = \mathbf{1}\}$. The set $\mathcal{X}$ is combinatorial and nonconvex, and the ratio structure in equation 19 and equation 20 further complicates the landscape, so the exact K-way problem is NP-hard in general. This motivates spectral relaxations and continuous surrogates followed by a discrete projection back onto $\mathcal{X}$.

B On the Suboptimality of Recursive Normalized Cut

Recursive partitioning applies a two-way Normalized Cut (Ncut) repeatedly until K parts are obtained. The procedure is simple and often efficient, yet it may be far from the global optimum of the K-way objective. This note explains why greedy bipartitions can deviate from the best K-way solution and highlights the mathematical and structural causes.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \boldsymbol{W})$ be an undirected graph with $|\mathcal{V}| = N$, affinity matrix $\boldsymbol{W} \in \mathbb{R}_+^{N\times N}$, and degree matrix $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{W}\mathbf{1})$. A K-way partition $\{\mathcal{A}_k\}_{k=1}^K$ minimizes

$$\mathrm{Ncut}(\{\mathcal{A}_k\}) = \sum_{k=1}^{K} \frac{\mathrm{cut}(\mathcal{A}_k, \mathcal{V}\backslash\mathcal{A}_k)}{\mathrm{vol}(\mathcal{A}_k)}, \qquad \mathrm{cut}(A,B) = \sum_{i\in A,\, j\in B} W_{ij}, \quad \mathrm{vol}(A) = \sum_{i\in A} d_i.$$

Exact minimization is NP-hard for $K > 2$. A common heuristic first finds a two-way Ncut $(\mathcal{A}, \mathcal{B})$ on $\mathcal{G}$, then recurses on the induced subgraphs until K clusters are formed. Despite its convenience, this scheme does not, in general, recover a globally optimal K-way partition.

Lack of optimal substructure. The K-way objective does not decompose into independent two-way subproblems. Once a boundary between $\mathcal{A}$ and $\mathcal{B}$ is fixed, subsequent decisions are constrained within induced subgraphs. A locally optimal bipartition $\mathrm{Ncut}(\mathcal{A}, \mathcal{B})$ need not be part of any globally optimal K-way solution.

Objective mismatch under recursion. Consider $K = 3$ with ground-truth communities A, B, C. If recursion first splits $S := A \cup B$ from C, then refines S into A and B, the recursive score equals

$$\underbrace{\frac{\mathrm{cut}(S,C)}{\mathrm{vol}(S)} + \frac{\mathrm{cut}(S,C)}{\mathrm{vol}(C)}}_{\text{first bipartition on } \mathcal{G}} + \underbrace{\frac{\mathrm{cut}_S(A,B)}{\mathrm{vol}_S(A)} + \frac{\mathrm{cut}_S(A,B)}{\mathrm{vol}_S(B)}}_{\text{second bipartition on the induced subgraph } S},$$

where cut_S and vol_S are computed in the induced subgraph on S. By contrast, the global 3-way objective is

$$\frac{\mathrm{cut}(A, B\cup C)}{\mathrm{vol}(A)} + \frac{\mathrm{cut}(B, A\cup C)}{\mathrm{vol}(B)} + \frac{\mathrm{cut}(C, A\cup B)}{\mathrm{vol}(C)}.$$

These expressions differ in the normalizers for the second stage: typically $\mathrm{vol}_S(A) < \mathrm{vol}(A)$ and $\mathrm{vol}_S(B) < \mathrm{vol}(B)$ whenever A or B has edges to C. Hence the recursive refinement overweights $\mathrm{cut}_S(A,B) = \mathrm{cut}(A,B)$ by dividing through smaller denominators, which can increase the total objective. There exist graphs where this "normalization gap" makes the recursive sum strictly larger than the global 3-way optimum, and the same phenomenon extends to $K > 3$.

Irreversibility and greedy commitment. Early boundaries are irrevocable. A bipartition that reduces the two-way score may merge distinct communities because the two-way objective favors grouping parts whose union has a large volume, even if a later K-way split would separate them. Once merged, separating them inside the induced subgraph ignores edges to the rest of the graph, which distorts normalization and may prevent recovery of the best global arrangement.

Spectral information loss in two-way splits. Two-way Ncut is driven by the Fiedler vector of $\boldsymbol{L}_{\mathrm{sym}}$. For $K > 2$, higher eigenvectors encode additional community structure. Sequentially applying a single eigenvector per split neglects joint information in the top K eigenvectors and can miss multi-community signals; rounding and recursion cannot, in general, reconstruct the simultaneous K-way structure.

Recursive bipartitioning is a useful heuristic, yet it imposes a sequential search on a global objective. Local optimality at each step is not sufficient for global optimality, especially on non-hierarchical graphs or when multiple spectral components are essential. Methods that optimize a K-way objective directly or allow global refinement can mitigate these failures.

C KURDYKA–ŁOJASIEWICZ CONVERGENCE ANALYSIS OF FALCON

Following the Kurdyka–Łojasiewicz (KL) Convergence analysis method in Jia et al. (2023), we establish the convergence of the entire sequence generated by Falcon to a stationary point of the objective function. This is achieved under the condition that an accumulation point of the sequence satisfies the Kurdyka–Łojasiewicz (KL) property.

C.1 PROBLEM, ALGORITHM, AND STANDING ASSUMPTIONS

Let $\mathbb{X} = \mathbb{R}^{N\times K}$ with Frobenius inner product and norm $\|\cdot\|$. We consider

$$\min_{\mathbf{X}\in\mathbb{X}} \ \psi(\mathbf{X}) := f(\mathbf{X}) + \phi(\mathbf{X}), \tag{21}$$

where $f : \mathbb{X} \to \mathbb{R}$ is C^1 and $\phi : \mathbb{X} \to (-\infty, +\infty]$ is proper, lower semicontinuous (lsc). In FALCON, $\phi = \iota_{\mathcal{V}_+}$ encodes the row-wise one-hot feasibility with nondegenerate cluster volumes; thus $\mathrm{dom}(\phi) = \mathcal{V}_+$.

In Falcon with monotone backtracking, given $\mathbf{X}^0 \in \mathrm{dom}(\phi)$ and parameters

$$\underline{\gamma} > 0, \qquad \eta > 1, \qquad \delta \in (0, 1),$$

for $k = 0, 1, 2, \ldots$ perform:

(A1) Backtracking loop. Initialize a trial stepsize $\gamma_k \in [\underline{\gamma}, +\infty)$.

(A2) Forward–backward subproblem. Compute

$$\mathbf{X}^{k+1} \in \arg\min_{\mathbf{Y}\in\mathbb{X}} \big\langle \nabla f(\mathbf{X}^k),\, \mathbf{Y} - \mathbf{X}^k \big\rangle + \frac{\gamma_k}{2}\|\mathbf{Y} - \mathbf{X}^k\|^2 + \phi(\mathbf{Y}). \tag{22}$$

When $\phi = \iota_{\mathcal{V}_+}$, equation 22 reduces to row-wise one-hot projection of $\mathbf{X}^k - \gamma_k^{-1}\nabla f(\mathbf{X}^k)$ onto $\mathcal{V}_+$.

(A3) Acceptance test. Accept if

$$\psi(\mathbf{X}^{k+1}) \le \psi(\mathbf{X}^k) - \frac{\delta\,\gamma_k}{2}\,\|\mathbf{X}^{k+1} - \mathbf{X}^k\|^2. \tag{23}$$

Otherwise set $\gamma_k \leftarrow \eta\,\gamma_k$ and repeat (A2)–(A3).

Assumption C.1 (Standing; cf. Assumption 3.2)**.**

(a) ψ is bounded from below on $\mathrm{dom}(\phi)$.

(b) ϕ is bounded from below by an affine function.

(c) ∇f is locally Lipschitz continuous.

We also assume the subproblem equation 22 is solvable for all k (true here since ϕ is an indicator and the objective in $\mathbf{Y}$ is strongly coercive).

Definition C.2 (KL property (at $\mathbf{X}^\star$))**.** A proper lsc ψ has the Kurdyka–Łojasiewicz (KL) property at $\mathbf{X}^\star$ if there exist $\eta > 0$, a neighborhood $\mathcal{U}$ of $\mathbf{X}^\star$, and a concave C^1 function $\chi : (0, \eta) \to \mathbb{R}_+$ with $\chi'(t) > 0$ such that

$$\chi'\big(\psi(\mathbf{X}) - \psi(\mathbf{X}^\star)\big) \cdot \operatorname{dist}\big(0, \partial\psi(\mathbf{X})\big) \;\geq\; 1$$

for all $\mathbf{X} \in \mathcal{U}$ with $\psi(\mathbf{X}^\star) < \psi(\mathbf{X}) < \psi(\mathbf{X}^\star) + \eta$. If $\chi(t) = c\, t^{1-\theta}$ with $\theta \in [0, 1)$, then θ is the KL exponent.

C.2 Preliminaries: Descent and Optimality Residual

Lemma C.3 (Finite termination of backtracking and sufficient decrease)**.** *Under Assumption C.1, the backtracking loop terminates after finitely many inner steps. The accepted $(\mathbf{X}^{k+1}, \gamma_k)$ satisfies equation 23. Consequently, $\{\psi(\mathbf{X}^k)\}$ is monotonically decreasing and bounded below; hence it converges.*

Proof. By local Lipschitz continuity, there exists $L_k < \infty$ with

$$f(\mathbf{Y}) \leq f(\mathbf{X}^k) + \langle \nabla f(\mathbf{X}^k), \mathbf{Y} - \mathbf{X}^k \rangle + \tfrac{L_k}{2} \|\mathbf{Y} - \mathbf{X}^k\|^2$$

for $\mathbf{Y}$ near $\mathbf{X}^k$. Combining this with the optimality of equation 22 yields equation 23 whenever $\gamma_k \geq L_k/(1-\delta)$. Since γ_k increases geometrically in backtracking, acceptance occurs in finitely many inner steps. The monotonicity and the lower bound of ψ (Assumption C.1(a)) imply convergence of $\psi(\mathbf{X}^k)$. □

Lemma C.4 (Optimality of the prox subproblem and residual bound)**.** *For the accepted $(\mathbf{X}^{k+1}, \gamma_k)$ one has*

$$0 \in \nabla f(\mathbf{X}^k) + \gamma_k(\mathbf{X}^{k+1} - \mathbf{X}^k) + \partial\phi(\mathbf{X}^{k+1}). \tag{24}$$

Hence

$$\operatorname{dist}\big(0, \partial\psi(\mathbf{X}^{k+1})\big) \;\leq\; \|\nabla f(\mathbf{X}^{k+1}) - \nabla f(\mathbf{X}^k)\| + \gamma_k \|\mathbf{X}^{k+1} - \mathbf{X}^k\|. \tag{25}$$

Proof. The inclusion equation 24 is the first-order condition of equation 22. Adding and subtracting $\nabla f(\mathbf{X}^{k+1})$ and using $\partial\psi(\mathbf{X}^{k+1}) = \nabla f(\mathbf{X}^{k+1}) + \partial\phi(\mathbf{X}^{k+1})$ gives equation 25. □

Step 1 (Lemma 4.1 analogue): Local Boundedness of Accepted Stepsizes

Lemma C.5 (Local boundedness of γ_k)**.** *Let $\mathbf{X}^\star$ be an accumulation point of $\{\mathbf{X}^k\}$. Then there exist a neighborhood $\mathcal{U}$ of $\mathbf{X}^\star$ and a constant $\overline{\gamma} < \infty$ such that, for all sufficiently large k with $\mathbf{X}^k \in \mathcal{U}$, the accepted stepsize satisfies $\gamma_k \leq \overline{\gamma}$.*

Proof. By Assumption C.1(c), ∇f is L_ρ-Lipschitz on a ball $\mathcal{U} = \{\mathbf{Z} : \|\mathbf{Z} - \mathbf{X}^\star\| \leq \rho\}$ for small $\rho > 0$. The descent argument in Lemma C.3 shows that equation 23 holds for any $\gamma \geq L_\rho/(1-\delta)$, provided $\mathbf{X}^k, \mathbf{X}^{k+1} \in \mathcal{U}'$ with slightly larger radius. Thus, once k is large so that iterates stay in $\mathcal{U}'$, the backtracking will accept a step with $\gamma_k \leq \overline{\gamma} := \eta\, L_\rho/(1-\delta)$. □

Step 2 (Lemma 4.2 analogue): Convergence of Function Values

Lemma C.6 ($\psi(\mathbf{X}^k) \to \psi(\mathbf{X}^\star)$)**.** *Let $\mathbf{X}^\star$ be an accumulation point of $\{\mathbf{X}^k\}$. Then $\psi(\mathbf{X}^k) \to \psi(\mathbf{X}^\star)$.*

Proof. By Lemma C.3, $\psi(\mathbf{X}^k)$ converges to some ψ^∞. Take $k_j \to \infty$ with $\mathbf{X}^{k_j} \to \mathbf{X}^\star$. Lower semicontinuity of ϕ and continuity of f give $\psi(\mathbf{X}^\star) \leq \liminf_j \psi(\mathbf{X}^{k_j}) = \psi^\infty$. Since $\psi(\mathbf{X}^k)$ is decreasing and $\mathbf{X}^\star$ lies in the same sublevel set, $\psi^\infty \leq \psi(\mathbf{X}^\star)$. Hence equality holds. □

STEP 3 (LEMMA 4.4 ANALOGUE): SUBGRADIENT–STEP RELATION

Lemma C.7 (Subgradient controlled by step length)**.** *There exist k_0 and $C > 0$ such that for all $k \geq k_0$,*

$$\operatorname{dist}\big(0, \partial\psi(\mathbf{X}^{k+1})\big) \;\leq\; C\,\|\mathbf{X}^{k+1} - \mathbf{X}^{k}\|. \tag{26}$$

Proof. Restrict to the tail where $\mathbf{X}^k$ lies in the neighborhood $\mathcal{U}$ of Lemma C.5. Then ∇f is L_ρ-Lipschitz and $\gamma_k \leq \overline{\gamma}$. From equation 25,

$$\operatorname{dist}\big(0, \partial\psi(\mathbf{X}^{k+1})\big) \leq (L_\rho + \overline{\gamma})\,\|\mathbf{X}^{k+1} - \mathbf{X}^{k}\|.$$

Set $C := L_\rho + \overline{\gamma}$ and choose k_0 large so that the neighborhood assumptions hold. □

C.3 MAIN THEOREM (THEOREM 4.5 ANALOGUE): GLOBAL CONVERGENCE

Theorem C.8 (Global convergence and finite length)**.** *Suppose Assumption C.1 holds and the sequence $\{\mathbf{X}^k\}$ generated by* FALCON *has an accumulation point $\mathbf{X}^\star$ at which ψ satisfies the KL property. Then $\{\mathbf{X}^k\}$ converges to $\mathbf{X}^\star$ and*

$$\sum_{k=0}^{\infty} \|\mathbf{X}^{k+1} - \mathbf{X}^{k}\| \;<\; \infty.$$

Proof. Let $\Delta_k := \psi(\mathbf{X}^k) - \psi(\mathbf{X}^\star) > 0$. By Lemma C.3 and Lemma C.6, $\Delta_k \downarrow 0$. By Lemma C.7, for $k \geq k_0$,

$$\operatorname{dist}\big(0, \partial\psi(\mathbf{X}^{k})\big) \;\leq\; C\,\|\mathbf{X}^{k} - \mathbf{X}^{k-1}\|.$$

The KL inequality at $\mathbf{X}^\star$ yields

$$\chi'(\Delta_k) \;\geq\; \frac{1}{C\,\|\mathbf{X}^k - \mathbf{X}^{k-1}\|} \qquad (k \geq k_0). \tag{27}$$

From the sufficient decrease equation 23 and $\gamma_k \geq \underline{\gamma}$,

$$\Delta_k - \Delta_{k+1} \;\geq\; \frac{\delta\,\underline{\gamma}}{2}\,\|\mathbf{X}^{k+1} - \mathbf{X}^{k}\|^2.$$

Concavity of χ implies

$$\chi(\Delta_k) - \chi(\Delta_{k+1}) \;\geq\; \chi'(\Delta_k)\,(\Delta_k - \Delta_{k+1}) \;\geq\; \frac{\delta\,\underline{\gamma}}{2C}\,\frac{\|\mathbf{X}^{k+1} - \mathbf{X}^{k}\|^2}{\|\mathbf{X}^k - \mathbf{X}^{k-1}\|}.$$

Summing from $k = k_0$ to m and telescoping the left-hand side gives

$$\sum_{k=k_0}^{m} \frac{\|\mathbf{X}^{k+1} - \mathbf{X}^{k}\|^2}{\|\mathbf{X}^k - \mathbf{X}^{k-1}\|} \;\leq\; \frac{2C}{\delta\,\underline{\gamma}}\big(\chi(\Delta_{k_0}) - \chi(\Delta_{m+1})\big) \;\leq\; \frac{2C}{\delta\,\underline{\gamma}}\,\chi(\Delta_{k_0}).$$

An elementary inequality then yields $\sum_{k\geq k_0} \|\mathbf{X}^{k+1} - \mathbf{X}^{k}\| < \infty$ (finite length). Hence $\{\mathbf{X}^k\}$ is Cauchy and converges; as $\mathbf{X}^\star$ is an accumulation point, the whole sequence converges to $\mathbf{X}^\star$. □

C.4 RATE (THEOREM 4.6 ANALOGUE): LINEAR CONVERGENCE FOR KL EXPONENT $\theta = \frac{1}{2}$

Theorem C.9 (Linear rates for $\theta = \frac{1}{2}$)**.** *Under the assumptions of Theorem C.8, suppose $\chi(t) = c\,t^{1/2}$ near $\mathbf{X}^\star$ (i.e., KL exponent $\theta = \frac{1}{2}$). Then there exist $q \in (0,1)$, $\omega > 0$, $\mu \in (0,1)$, and k_1 such that*

$$\begin{aligned}
&\text{(Q-linear values)} && \psi(\mathbf{X}^{k+1}) - \psi(\mathbf{X}^\star) \;\leq\; q\big(\psi(\mathbf{X}^k) - \psi(\mathbf{X}^\star)\big), \quad \forall k \geq k_1,\\
&\text{(R-linear iterates)} && \|\mathbf{X}^k - \mathbf{X}^\star\| \;\leq\; \omega\,\mu^k, \quad \forall k \geq k_1.
\end{aligned}$$

Proof. With $\chi(t) = c\,t^{1/2}$, equation 27 becomes

$$\frac{c}{2}\,(\Delta_k)^{-1/2} \;\geq\; \frac{1}{C\,\|\mathbf{X}^k - \mathbf{X}^{k-1}\|}.$$

Combining with $\Delta_k - \Delta_{k+1} \geq \frac{\delta\gamma}{2}\|\mathbf{X}^{k+1} - \mathbf{X}^k\|^2$ and following the standard argument in the proof of Theorem 4.6 (cf. Section 4), one obtains a linear contraction $\Delta_{k+1} \leq q\,\Delta_k$ for all large k and some $q \in (0,1)$. The R-linear rate for iterates follows from summing the sufficient decrease and bounding $\|\mathbf{X}^{k+1} - \mathbf{X}^k\|^2$ in terms of $\Delta_k - \Delta_{k+1}$ to dominate the tail distance to $\mathbf{X}^\star$ by a geometric series. □

C.5 Remarks Specific to Falcon

Feasible set and KL property. When $\phi = \iota_{\mathcal{V}_+}$ and f is analytic in a neighborhood of a limit point $\mathbf{X}^\star \in \mathcal{V}_+$ with nondegenerate cluster volumes, $\psi = f + \iota_{\mathcal{V}_+}$ is definable in an o-minimal structure, thus satisfies the KL property at $\mathbf{X}^\star$.

Local Lipschitz of ∇f in token-NCut. For the normalized association objective used by Falcon, the explicit gradient formula shows ∇f is locally Lipschitz on any region where per-cluster volumes are bounded away from zero, which is enforced by $\mathrm{dom}(\phi) = \mathcal{V}_+$. This verifies Assumption C.1(c) near accumulation points.

Template note. The structure of Lemmas C.5–C.7 and Theorems C.8–C.9 follows the Section 4 framework (descent, subgradient bound, KL) of Jia et al. (2023).

D Time Complexity of the Falcon

We analyze the proximal step of Falcon under an affinity $\boldsymbol{W} \in \mathbb{R}^{N\times N}$. Let $\boldsymbol{X} \in \{0,1\}^{N\times K}$ be row-one-hot. In one accepted outer iteration, the algorithm forms

$$\boldsymbol{S} = \boldsymbol{W}\boldsymbol{X}, \qquad \boldsymbol{q} = (\boldsymbol{X} \odot \boldsymbol{S})^\top \mathbf{1}, \qquad \boldsymbol{v} = \boldsymbol{X}^\top \boldsymbol{d}, \qquad \boldsymbol{d} = \boldsymbol{W}\mathbf{1}, \qquad \boldsymbol{D} = \mathrm{diag}(\boldsymbol{d}),$$

computes the gradient

$$\nabla f(\boldsymbol{X}) = 2\big(\boldsymbol{S}\,\boldsymbol{v}^{-T} - (\boldsymbol{D}\boldsymbol{X})\,\boldsymbol{c}^\top\big), \qquad c_k = q_k / v_k^2,$$

performs the row-wise $\arg\max$ update, and evaluates the Armijo-type acceptance criterion. The dominant costs are:

$$\begin{aligned}\text{Per iteration cost} \;=\; &\underbrace{\boldsymbol{S} = \boldsymbol{W}\boldsymbol{X}}_{-\,\mathcal{O}(N^2K)} \\ &+ \underbrace{\text{column scalings and subtraction in } \nabla f}_{-\,\mathcal{O}(NK)} \\ &+ \underbrace{\text{row-wise } \arg\max}_{-\,\mathcal{O}(NK)} \\ &+ \underbrace{\boldsymbol{S}^+ = \boldsymbol{W}\boldsymbol{X}^+}_{-\,\mathcal{O}(N^2K)}.\end{aligned}$$

Hence one accepted iteration costs is $\mathcal{O}(N^2K)$ (lower-order $\mathcal{O}(NK)$ omitted).

If $B_t \geq 1$ denotes the number of line-search trials in iteration t (including the accepted one), each trial repeats $\boldsymbol{W}\boldsymbol{X}^+$ and light $\mathcal{O}(NK)$ work, so the cost of iteration t is $\mathcal{O}(B_t\,N^2K)$. Near the limit set, Armijo acceptance holds for any $\tau \geq L_\rho/(1-\delta)$; starting from $\tau_0 > 0$ and multiplying by $\gamma > 1$ gives $B_t \;\leq\; 1 + \left\lceil \log_\gamma\!\left(\frac{L_\rho}{(1-\delta)\tau_0}\right)\right\rceil$, i.e., a uniform *constant* bound, so backtracking is only a small multiplicative factor. If $B_t = \mathcal{O}(1)$, with T accepted outer iterations,

$$\text{Total cost} \;=\; \mathcal{O}\Big(\big(\textstyle\sum_{t=1}^{T} B_t\big)\,N^2K\Big) \;=\; \mathcal{O}(T\,N^2K) \qquad T, K \ll N$$

Classical spectral methods (eigenvector computation for the Laplacian or normalized affinity) and recursive NCut typically require solving an $N \times N$ eigenproblem, leading to cubic $\mathcal{O}(N^3)$ time in standard dense linear algebra. In contrast, Falcon's core cost is quadratic $\mathcal{O}(N^2)$ for constant K and T. Moreover, our KL-based analysis guarantees convergence of the whole sequence (and linear rates when the KL exponent is $1/2$), a property that plain power or eigen-solvers used in recursive NCut do not directly provide at the discrete assignment level.

E ROPERTIES AND TIME COMPLEXITY OF PAMR AND NAMR

Let $\Omega \subset \mathbb{Z}^2$ be the pixel grid, $|\Omega| = N$, with image $I : \Omega \to \mathbb{R}^3$. The initial segmentation is $M_0 \in [0,1]^{(C+1)\times N}$ (one-vs-rest probabilities from NN-upsampled token masks). For a set of dilations $D \subset \mathbb{N}^+$, define displacement sets Δ_d of the 8 neighbors at distance d and $\Delta = \bigcup_{d\in D} \Delta_d$ with $P = 8|D|$. For $\delta \in \Delta$, the shift operator $(S_\delta x)(i) = x(i+\delta)$ uses border replication.

For each (i, δ), define channelwise discrepancies

$$r_k(i,\delta) = \big|I_k(i) - I_k(i+\delta)\big|, \quad \bar{r}(i,\delta) = \tfrac{1}{3}\sum_{k=1}^{3} r_k(i,\delta), \quad \sigma(i) = \mathrm{Std}_{\delta\in\Delta}\,\bar{r}(i,\delta).$$

NAMR replaces $\bar{r}$ by $\bar{r}^{\mathrm{nl}}(i,\delta) = \frac{1}{3}\sum_k \big|\phi(I_k(i) - I_k(i+\delta))\big|$ with a fixed pointwise ϕ.

PAMR as anisotropic diffusion on a row-stochastic graph. For a stabilization constant $\varepsilon > 0$ and (optionally) a scale factor $\lambda > 0$ absorbed into σ, define directional logits and softmax weights

$$\ell(i,\delta) = -\frac{\bar{r}(i,\delta)}{\varepsilon + \sigma(i)}, \qquad p(i,\delta) = \frac{\exp\{\ell(i,\delta)\}}{\sum_{\delta'\in\Delta}\exp\{\ell(i,\delta')\}}.$$

Collect $p(i,\delta)$ into a sparse row-stochastic matrix $W \in \mathbb{R}^{N\times N}$ with $W(i, i+\delta) = p(i,\delta)$ and zeros elsewhere. For each class c,

$$m_{t+1}^{(c)} = W\, m_t^{(c)}, \qquad t = 0,\dots,T-1, \qquad M^0 := M_0, \quad M^T = [m_T^{(1)} \ \cdots \ m_T^{(C+1)}].$$

Stability. Since W is row-stochastic, each update is a convex combination, $\|m_{t+1}^{(c)}\|_\infty \le \|m_t^{(c)}\|_\infty$.
Energy view (fixed W). A finite number of Jacobi steps approximates minimizing

$$E(M) = \tfrac{1}{2}\sum_{i,j} W(i,j)\,\|M(:,i) - M(:,j)\|_2^2 \ + \ \alpha\,\|M - M_0\|_2^2,$$

with an optional data-fidelity $\alpha \ge 0$. Early stopping ($T \approx$ 5–10) balances denoising and edge preservation.

NAMR: non-linear contrast and multi-temperature aggregation. For $\mathcal{T} = \{\tau_0, \dots, \tau_{m-1}\}$,

$$\ell^{(\tau)}(i,\delta) = -\frac{\bar{r}^{\mathrm{nl}}(i,\delta)}{\varepsilon + \tau\,\sigma(i)}, \quad p^{(\tau)}(i,\delta) = \mathrm{softmax}_{\delta\in\Delta}\,\ell^{(\tau)}(i,\delta), \quad W^{(\tau)}(i, i+\delta) = p^{(\tau)}(i,\delta).$$

Run T steps per temperature,

$$M_T^{(\tau)} = (W^{(\tau)})^T M_0, \qquad \bar{M}_T = \tfrac{1}{m}\sum_{\tau\in\mathcal{T}} M_T^{(\tau)}, \qquad \hat{y}(i) = \arg\max_c \bar{M}_T^{(c)}(i).$$

Small τ downweights cross-edge transport; large τ encourages within-region smoothing. The average $\bar{M}_T$ behaves like a multi-bandwidth anisotropic diffusion, improving robustness across appearances.

Time Complexity. Let $N = HW$, classes $C{+}1$, directions $P = 8|D|$, iterations T, and $m = |\mathcal{T}|$ (for NAMR). Constructing $\bar{r}$ and σ costs $\mathcal{O}(N\cdot 3\cdot P)$. Each message-passing step costs $\mathcal{O}(N\cdot C\cdot P)$. Hence:

$$\text{PAMR: } \mathcal{O}\big(N(3P + TCP)\big), \qquad \text{NAMR: } \mathcal{O}\big(N(3P + mTCP)\big).$$

The framework is drop-in: RGB can be replaced by any guidance features (e.g., depth encoders) without changing derivations.

Figure 5: **Qualitative segmentation results of Falcon across diverse domains.** Examples span natural scenes, urban landscapes, medical CT, paintings, cartoons, and planetary imagery. Falcon consistently produces coherent partitions with sharp boundaries and stable region assignment, demonstrating strong generalization beyond a single modality.

F FALCON SEGMENTATION RESULTS VISUALIZATION

G Evaluation Runtime Results

Table 3 details the decomposition of runtime into *Data Preparation*, *Feature Extraction*, and *Mask Generation*. Falcon outperforms baselines across all 18 dataset–encoder pairs. Representative examples include: (i) *DINOv3-B, Cityscapes*: Total Time drops from 784.04s (DiffCut) to 87.47s ($\sim$8.9$\times$), and Mask Generation from 747.97s to 52.49s ($\sim$14.3$\times$). (ii) *SD2.1, Context*: Total Time reduces from 2535.03s to 1173.87s ($\sim$2.2$\times$), and Mask Generation from 1546.75s to 182.86s ($\sim$8.5$\times$). (iii) *SSD-1B, Cityscapes*: even when preprocessing dominates, Falcon still lowers Total Time (177.27s $\rightarrow$ 114.67s; $\sim$1.5$\times$) and shrinks Mask Generation (78.11s $\rightarrow$ 15.46s; $\sim$5.0$\times$).

Table 3: Total evaluation time on various datasets. Best **Total time** (s) and **Mask Generation** (s) per dataset are bold.

Encoder	Benchmarks	Method	Total Time	Data Preparation	Feature Ext.	Mask Generation
SSD-1B	Cityscapes	AutoSC	126.63	23.02	76.25	27.36
		DiffCut	177.27	22.97	76.19	78.11
		Falcon	**114.67**	22.96	76.25	**15.46**
	VOC	AutoSC	368.96	68.61	220.71	79.64
		DiffCut	483.05	71.21	220.39	191.45
		Falcon	**335.31**	68.88	220.96	**45.47**
	Context	AutoSC	1288.12	232.30	775.63	280.19
		DiffCut	1700.30	237.37	775.37	687.56
		Falcon	**1170.80**	233.73	775.74	**161.33**
	COCO-Stuff	AutoSC	551.18	100.89	330.78	119.51
		DiffCut	734.07	99.92	330.73	303.42
		Falcon	**497.55**	98.84	330.91	**67.80**
	COCO-Object	AutoSC	556.11	105.84	330.72	119.55
		DiffCut	748.96	114.85	330.68	303.43
		Falcon	**506.54**	106.48	331.23	**68.83**
	ADE20K	AutoSC	505.73	91.34	304.17	110.22
		DiffCut	686.08	90.65	304.03	291.40
		Falcon	**459.25**	92.13	304.77	**62.35**
SD2.1	Cityscapes	AutoSC	129.15	23.13	74.44	31.58
		DiffCut	410.24	23.12	74.27	312.85
		Falcon	**114.35**	23.13	74.58	**16.64**
	VOC	AutoSC	374.69	68.68	214.97	91.04
		DiffCut	719.16	68.37	214.77	436.02
		Falcon	**335.38**	69.51	214.62	**51.25**
	Context	AutoSC	1313.00	234.68	756.28	322.04
		DiffCut	2535.03	233.09	755.19	1546.75
		Falcon	**1173.87**	234.24	756.77	**182.86**
	COCO-Stuff	AutoSC	561.70	101.46	322.60	137.64
		DiffCut	1197.13	99.31	322.27	775.55
		Falcon	**498.58**	97.88	322.90	**77.80**
	COCO-Object	AutoSC	566.79	106.95	322.52	137.32
		DiffCut	1208.45	103.69	322.18	782.58
		Falcon	**503.54**	102.28	323.37	**77.89**
	ADE20K	AutoSC	514.99	91.26	296.62	127.11
		DiffCut	1090.35	90.80	296.25	703.30
		Falcon	**457.40**	89.28	296.60	**71.52**
DINOv3-b	Cityscapes	AutoSC	242.67	23.07	12.14	207.46
		DiffCut	784.04	23.93	12.14	747.97
		Falcon	**87.47**	22.48	12.50	**52.49**
	VOC	AutoSC	718.39	69.10	35.00	614.29
		DiffCut	1950.92	66.33	34.97	1849.62
		Falcon	**238.46**	66.34	35.12	**137.00**
	Context	AutoSC	2528.36	236.62	122.93	2168.81
		DiffCut	6935.53	234.54	122.85	6578.14
		Falcon	**903.28**	237.21	123.24	**542.83**
	COCO-Stuff	AutoSC	1073.20	100.08	52.47	920.65
		DiffCut	2916.03	99.05	52.43	2764.55
		Falcon	**384.14**	101.16	52.41	**230.57**
	COCO-Object	AutoSC	1184.79	107.20	52.47	1025.12
		DiffCut	2947.12	102.48	52.43	2792.21
		Falcon	**386.91**	102.80	52.72	**231.39**
	ADE20K	AutoSC	996.57	93.55	48.26	854.76
		DiffCut	2787.44	90.47	48.22	2648.75
		Falcon	**351.39**	90.10	48.46	**212.83**