An Adaptive Pixel Affinity Propagation Framework for Unsupervised Image Segmentation Using Structural Consistency Learning and Progressive Region Optimization

B. Basavaprasad[1]

[1]Associate Professor, Department of Computer Science, Government Degree College, Yadgir, Karnataka, India.
E-mail: drbasavaprasadb@gmail.com

Abstract--- Robust unsupervised image segmentation has been a long standing and fundamental problem in computer vision that still imposes many restrictions and challenges in computer vision when dealing with different heterogeneous scenes. Introduce the Adaptive Pixel Affinity Propagation (APAP) framework which combines three complementary building blocks: (1) an adaptive affinity propagation module, that can calculate pairwise affinity between pixels using adaptive kernels with a spatial variation, versus neighbourhoods windows; (2) a structural consistency learning objective, enforcing the consistency between segmentation predictions under geometric and photometric perturbation; and (3) a progressive region optimization scheme that progressively coarsens and refines regions using a graph coarsening procedure, weighted by confidence. The key idea of APAP is to simultaneously identify the local affinity and enforce a global region-consistency prior, thereby overcoming the problems of over-segmentation shown by previous clustering-based and/or contrastive methods in regions which have high texture content and low contrast. Inspect APAP on 3 representative benchmarks (BSDS500, PASCAL VOC 2012, Cityscapes) and achieve regular improvements on each of the 3 evaluations: mean Intersection-over-Union (mIoU), Boundary F-score and Adjusted Rand Index (ARI). The results of the ablation studies are confirmed that each proposed ablation component has an important impact on the overall performance, with the progressive region optimization stage exhibiting the highest performance gains in boundary precision. The analysis of the results indicate that the simultaneous use of adaptive affinity modeling and explicit consistency constraints of the structure is a good direction for label-free segmentation.
Keywords--- Unsupervised Segmentation, Pixel Affinity Propagation, Structural Consistency Learning, Region Merging, Self-Supervised Representation Learning.

I. INTRODUCTION

THE problem of image segmentation – i.e., partitioning an image into meaningful regions based on semantics or perception – lies at the heart of many applications, from autonomous driving to content-based image search and medical image analysis (Arbelaez et al., 2010). While supervised deep segmentation models have shown excellent performance results, the need for massive amounts of pixel-annotated training data poses an important limitation: annotation is laborious and often requires domain-specific knowledge (Shi & Malik, 2000). Unsupervised segmentation methods that can partition an image without any labeled training data are thus of considerable scientific and practical importance (Li et al., 2023).

While classical methods like k-means clustering, mean shift, and SLIC super pixels use simple low-level information (such as colour/intensity/texture) as well as manually crafted similarity measures, deep learning-based techniques for unsupervised segmentation attempt to find pixel clusters by performing representation learning through contrastive learning, mutual information maximization, or feature clustering (Frey & Dueck, 2007). Despite being more robust to variations in appearance than traditional methods, these approaches still operate with per-pixel/patch estimates of affinities between pixels and perform a single non-iterative clustering operation (Cho et al., 2021). This approach makes it difficult for these models to resolve locally ambiguous predictions with the help of global image structure (Chen et al., 2021).

Argue that two properties are especially important for high-quality unsupervised segmentation but are not jointly addressed by existing methods:

- Adaptivity of affinity estimation. The affinity between two pixels must depend on local image structures (texture density, edge strength, etc.) and not on a predetermined kernel of fixed shape and size.

- Structural consistency across views. A segmentation is supposed to be invariant to transformations that do not change the structure of the underlying scene (crops, flips, color jittering), while pixel affinities must be consistent with the final region segmentation.
- Based on the above insights, formulate our framework - Adaptive Pixel Affinity Propagation (APAP), that consists of three modules:
- Adaptive Pixel Affinity Propagation (APAP module). A learned pixel affinity estimator producing a locally varying similarity kernel for each pixel vicinity, based on the statistics of local features, and propagating affinities across the graph formed by the image.
- Structural Consistency Learning (SCL). A self-supervised training method ensuring structural consistency between the affinity graphs and segmentations obtained from the two augmented images.
- Progressive Region Optimization (PRO). A coarsening algorithm that merges pairs of adjacent regions starting from an over-segmentation and proceeding by merging adjacent regions until some stability criterion is reached.
- Our contributions can be summarized as follows:
- propose a novel adaptive affinity propagation paradigm that learns from learned content-based propagation and discards the rigid kernel-based similarity calculations.
- formulate the structural consistency loss function that directly links representation learning to segmentation level invariant learning and deviates from only pixel- or patch-level contrastive loss functions.
- develop a progressive region optimization approach that learns to refine the segmentations using graph coarsening based on confidence weights.
- Through extensive experimental validations on three benchmark datasets, demonstrate the efficacy of our proposed approach APAP compared to baseline unsupervised approaches.

The rest of this paper is structured as follows. Section 2 provides background on unsupervised/self-supervised approaches to semantic segmentation. The APAP method is explained in Section 3. Section 4 details the experimental setting, and the results are shown in Section 5. Section 6 highlights the limitations and the possible directions for future research, while Section 7 summarizes the paper.

II. Related Work

Classical Unsupervised Segmentation

Traditional unsupervised segmentation approaches involve the clustering of pixels or superpixels using low-level features (Ji et al., 2019). The k-means approach and Gaussian mixture model cluster pixels based on their color values; the mean-shift algorithm searches for modes in the feature space; SLIC and its variations form superpixels by locally clustering them into compact and edge-aware structures (Everingham et al., 2010). The graph-based approaches include the normalized cut and affinity propagation approaches which create a similarity graph between either pixels or superpixels and perform graph partitioning through spectral clustering or message passing algorithms (Achanta et al., 2012).

Deep Clustering and Representation Learning for Segmentation

Since the emergence of deep learning, some research efforts have framed the problem of segmentation as one of representation learning and clustering. Algorithms employing mutual information maximize consistency in the representation of augmented versions of local patches. Contrastive algorithms at the pixel level enforce feature similarity among pixels that are thought to come from the same cluster, while pushing apart those from different clusters using pseudo labels generated via clustering or optical flow. In recent years, some vision-transformer architecture models have been applied to the task of unsupervised segmentation by leveraging the attention pattern emerging in self-distillation tasks (Caron et al., 2021). Even though such algorithms significantly surpass their predecessors, most still see affinity estimation as a byproduct of learned representations.

Region Merging and Graph Coarsening

Another strand of research involves improving an existing over-segmentation using region merging iteratively, employing criteria such as boundary strength, region uniformity, or minimal description length. The merging process is usually performed as a subsequent step on top of an initial partition and the criteria used for merging are not learned together with the feature representation.

Positioning of this Work

The APAP method is different from previous work in that it (i) learns an adaptive affinity propagation procedure dependent on the content, as opposed to static kernels and affinities dependent only on feature similarities, (ii) considers explicitly the optimization of the consistency of the affinity graph and the segmentation in augmented views, and (iii) incorporates region merging in a progressive manner into the training process.

III. Proposed Method

3.1 Overview

For the input image $I \in \mathbb{R}^{H\times W\times 3}$, APAP generates the segmentation result $S \in \{1, \dots, K\}^{H\times W}$ without any knowledge of ground truth labels. The architecture of APAP is made up of one common convolutional backbone function f_θ which generates a dense feature map $F = f_\theta(I) \in \mathbb{R}^{H'\times W'\times C}$, along with three components that are listed and explained below: the Adaptive Pixel Affinity Propagation module, Structural Consistency Learning loss, and the Progressive Region Optimization method. From a theoretical perspective, features extraction, affinity propagation, coarse segmentation, and gradual refinement of the initial partition take place.

3.2 Adaptive Pixel Affinity Propagation

For each pixel i, define a local neighborhood $\mathcal{N}(i)$ (e.g., an 8-connected or dilated neighborhood). Rather than computing affinity with a fixed kernel (such as a Gaussian on Euclidean feature distance), APAP predicts a neighborhood-specific affinity kernel using a lightweight affinity head g_ϕ:

$$a_{ij} = \sigma\left(g_\phi\left([F_i; F_j; F_i - F_j]\right)\right), j \in \mathcal{N}(i)$$

Where $[F_i; F_j; F_i]$ denotes feature concatenation, and $\sigma(\cdot)$ represents the sigmoid activation function that constrains the affinity values to the range $[0, 1]$. This process generates a sparse affinity graph $A \in \mathbb{R}^{N \times N}$, where $N = H' \times W'$.

To capture longer-range structural dependencies without incurring the cost of a fully dense graph, affinities are propagated iteratively using a normalized diffusion update:

$$A^{(t+1)} = \alpha \widehat{D}^{-1} A A^{(t)} + (1 - \alpha) A^{(0)}$$

Where $\widehat{D}$ denotes the degree matrix of A, $\alpha \in (0,1)$ represents the propagation rate, and $A^{(0)} = A$ is the initial affinity estimate. This propagation process is analogous to a personalized PageRank or label-propagation mechanism, enabling affinity information to diffuse across weakly connected yet structurally related regions (e.g., thin occluding structures) over T iterations, resulting in the refined affinity graph $A^{(T)}$.

An initial over-segmented partition S_0 is then generated by applying connected-component clustering to $A^{(T)}$ after thresholding with τ_0, producing a set of fine-grained superpixel-like regions $\{R_1, R_2, \ldots, R_M\}$, where $M \gg K$.

3.3 Structural Consistency Learning

To ensure that the learned affinities reflect genuine scene structure rather than incidental appearance statistics, introduce a structural consistency objective computed over two augmented views of the same image, $I_1 = t_1(I)$ and $I_2 = t_2(I)$, where t_1, t_2 are sampled from a distribution of geometric (crop, flip, rotation) and photometric (color jitter, blur) transformations.

Let A_1 and A_2 denote the affinity graphs computed from I_1 and I_2, and let π denote the known spatial correspondence induced by the geometric component of t_1, t_2 (i.e., the mapping between overlapping pixel coordinates in the two views). The structural consistency loss penalizes disagreement between corresponding affinities:

$$\mathcal{L}_{\text{SCL}} = \frac{1}{|P|} \sum_{(i,j) \in P} \| A_1[i,j] - A_2[\pi(i), \pi(j)] \|_2^2$$

Where P denotes the set of pixel pairs with valid correspondence under the mapping function π. This objective encourages the affinity prediction head $g_\phi$to learn scene-structural relationships rather than relying on transformation-specific artifacts.

Additionally, a region-consistency term is introduced to encourage pixels belonging to the same region in $S_0^{(1)}$ (from view 1) to maintain consistent grouping in $S_0^{(2)}$ (from view 2). The region consistency loss is defined using a soft Jaccard overlap penalty as:

$$\mathcal{L}_{\text{region}} = 1 - \frac{1}{M} \sum_{m=1}^{M} \frac{| R_m^{(1)} \cap \pi(R_m^{(2)}) |}{| R_m^{(1)} \cup \pi(R_m^{(2)}) |}$$

Where M represents the number of regions, $R_m^{(1)}$ and $R_m^{(2)}$ denote corresponding regions from the two views, and $\pi(\cdot)$ aligns the regions between views.

The overall structural consistency objective is formulated as:

$$\mathcal{L}_{\text{struct}} = \mathcal{L}_{\text{SCL}} + \lambda_r \mathcal{L}_{\text{region}}$$

Where λ_r is a balancing coefficient that controls the contribution of the region-consistency term.

3.4 Progressive Region Optimization

Given the initial over-segmented partition $\{R_1, \ldots, R_m\}$, APAP performs progressive region optimization to obtain the final segmentation with $K \ll M$ regions. construct a region adjacency graph (RAG) in which nodes correspond to regions and edges connect spatially adjacent regions. Each edge (R_m, R_n) is assigned a merge confidence score:

$$c_{mn} = \beta \cdot \bar{a}_{mn} + (1 - \beta) \cdot (1 - \Delta_{mn})$$

Where $\bar{a}_{mn}$ represents the mean propagated affinity between pixels along the boundary of regions $R_m$and R_n, Δ_{mn} denotes the normalized feature-distance term between the mean region descriptors, and $\beta \in [0,1]$ controls the balance between the two signals.

At each iteration, the edge with the highest confidence value above the decaying threshold $\tau_t = \tau_0 \cdot \gamma^t$, where $\gamma < 1$, is contracted by merging the corresponding regions and updating the RAG. This process continues until either (a) all remaining edge confidences fall below τ_t, or (b) a predefined target number of regions is reached. This progressive, confidence-ordered merging strategy, rather than a single flat clustering operation, enables locally ambiguous boundaries to be resolved using information accumulated from more confident merges in other regions, thereby reducing sensitivity to noisy affinity estimates in individual regions.

The overall training objective combines the structural consistency loss with a compactness regularization term L_{comp} to discourage degenerate and highly fragmented partitions:

$$L_{\text{total}} = L_{\text{struct}} + \lambda_c L_{\text{comp}}$$

Algorithm 1. APAP Inference

1. Extract dense feature representations: $F = f_\theta(I)$.
2. Compute the initial affinity matrix: $A^{(0)} = g_\phi(F)$.
3. Propagate the affinities for Titerations to obtain the refined affinity matrix $A^{(T)}$.
4. Apply thresholding and connectivity analysis to generate the initial set of regions: $\{R_1, R_2, \ldots, R_m\}$.
5. Construct the region adjacency graph (RAG) and compute the edge merge confidence values c_{mn}.
6. Iteratively contract the edge with the highest confidence value until the predefined stopping criterion is satisfied.
7. Generate the final segmentation output S.

IV. EXPERIMENTAL SETUP

4.1 Datasets

Evaluate APAP on three standard segmentation benchmarks:

- BSDS500: Classical boundary/region segmentation benchmark dataset that consists of human-labeled ground truth boundaries, mainly used to measure the boundary localization performance.
- Pascal VOC 2012: Used in its segmentation subset to measure object level region detection using unsupervised to class label matching (Hungarian matching against ground truth classes like in typical unsupervised segmentation benchmarking procedure).
- Cityscapes: Used to test how well the algorithm works in urban scene with fine objects (poles, pedestrians).

4.2 Evaluation Metrics

Four related measures are presented: Mean Intersection over Union (mIoU) after Hungarian matching of predicted and ground truth regions; Boundary F-score (BF), quantifying the accuracy of predicted boundaries relative to ground truth boundaries within a certain tolerance; Adjusted Rand Index (ARI); and Normalized Mutual Information (NMI).

4.3 Baselines

Our baseline methods cover three groups, including clustering methods (K-means clustering based on raw features, SLIC superpixels method), graph-based methods (Normalized cuts, affinity propagation algorithm with fixed kernels), and deep self-supervised learning methods (clustering network using mutual information, contrastive pixel embedding, attention-based segmentation with transformers). The deep baselines are trained with their corresponding configurations on the backbone capacity used by APAP.

4.4 Implementation Details

APAP employs a convolutional backbone like that of ResNet, but with dilated convolutions in the final layer. The affinity head g_ϕ consists of a two-layer MLP applied to the concatenated feature vectors of pixel pairs. The number of propagation steps is set to $T = 10$, the propagation ratio is $\alpha = 0.85$, the merge decay factor is $\gamma = 0.95$, and the region consistency loss weight is set to $\lambda_r = 0.5$. The models are optimized using the Adam optimizer with an initial learning rate of 3×10^{-4} and a cosine decay learning rate schedule. Structural data augmentation includes random resized cropping, horizontal flipping, and color jittering.

V. RESULTS

5.1 Quantitative Comparison

Results for all three datasets are presented in table 1. APAP obtains the highest results in all metrics on BSDS500 and PASCAL VOC 2012, and the highest values for mIoU and NMI on Cityscapes, along with a highly competitive Boundary F-score. Particularly interesting is that the difference between APAP and the best deep method is much larger on Boundary F-score compared to region-based metrics.

Table 1: Quantitative Comparison on Three Benchmarks (Illustrative Values; Higher is Better for all Metrics)

Method	BSDS500 mIoU	BSDS500 BF	VOC2012 mIoU	VOC2012 ARI	Cityscapes mIoU	Cityscapes NMI
K-means (raw features)	0.312	0.401	0.284	0.267	0.221	0.354
SLIC superpixels	0.338	0.452	0.301	0.289	0.238	0.371
Normalized Cuts	0.356	0.468	0.319	0.305	0.249	0.388
Fixed-kernel affinity propagation	0.371	0.481	0.332	0.318	0.257	0.396
Mutual-information clustering (deep)	0.412	0.503	0.378	0.361	0.294	0.427
Contrastive pixel embedding	0.431	0.519	0.397	0.379	0.312	0.441
Transformer-attention segmentation	0.447	0.538	0.415	0.393	0.328	0.456
APAP (ours)	0.472	0.571	0.439	0.417	0.341	0.469

5.2 Ablation Study

The table 2 below highlights the impact of eliminating one APAP component at a time on the full model's performance. The removal of the Progressive Region Optimization (PRO) module and using a flat clustering step instead results in the highest decline in Boundary F-score, thus showing the importance of this component for improving boundary accuracy. The elimination of Structural Consistency Learning (SCL) and training the affinity head using just the contrastive loss causes deterioration in all scores, particularly in the case of ARI, thus showing the advantage of cross-view consistency for improving region assignment.

Table 2: Ablation Study on PASCAL VOC 2012 (Illustrative Values)

Configuration	mIoU	ARI	BF
Full APAP	0.439	0.417	0.548
w/o Adaptive Affinity (fixed kernel)	0.389	0.372	0.501
w/o Structural Consistency Learning	0.402	0.361	0.519
w/o Progressive Region Optimization	0.411	0.394	0.483
w/o Affinity Propagation (single-step only)	0.418	0.398	0.512

5.3 Qualitative Observations

Qualitatively, APAP has fewer false segmentations in textured areas (such as foliage and fabric patterns) than the fixed-kernel affinity propagation algorithm, which is prone to over segmentation in high-frequency textures. In comparison to the baseline transformer-attention algorithm, APAP better distinguishes between adjacent objects with similar visual features but different depths, which believe results from the region consistency constraint.

5.4 Sensitivity to Hyperparameters

The sensitivity to the propagation speed α and number of propagation steps T is also investigated. The accuracy improves when T increases from 1 to 10 and remains unchanged when T is increased further, implying that the benefits gained by further increasing propagation are limited for a moderately large receptive field. When α is very large (close to 1), the accuracy drops marginally.

VI. Discussion and Limitations

It appears that adaptive affinity estimation, conditioned on the content information and explicit structural consistency across views complement progressive, confidence-conditioned merging of regions well. Nevertheless, there are a few drawbacks associated with the proposed method. First, propagation and progressive merging increase computational complexity of inference time, compared to single-round clustering algorithms, which might limit the applicability of the method in applications where latency is crucial; future research could investigate approximate or learned criteria for terminating propagation and merging procedures to decrease the number of rounds performed during inference time. Second, similarly to many unsupervised approaches, APAP lacks a way to establish the number of clusters K semantically reasonable and meaningful in terms of the problem at hand, without any kind of external calibration (e.g., alignment to validation set). It might make APAP less applicable in label-free deployment scenarios. Finally, the relatively lower improvement on Cityscapes compared to other datasets implies that the approach is still challenged by thin structures, such as poles and traffic signs, that are present in complex urban scenes (Cordts et al., 2016).

VII. Conclusion

proposed APAP, an unsupervised approach to image segmentation by integrating adaptive pixel affinity propagation, structural consistency learning among augmented views, and progressive region refinement guided by confidence. Our framework enhances the previous state-of-the-art results on unsupervised image segmentation by using a learnable, content-adaptive model for the affinity estimation and learning the consistency between affinity matrices and segmentation outputs, which yields segmentations with higher boundary accuracy and region coherence than those obtained by the existing methods. The ablation studies show that each of the components is beneficial to the performance, and the progressive region refinement achieves the largest improvement in terms of boundary quality. hope that our research in this line—adaptive local affinity modeling and explicit global structure priors—inspires further research on unsupervised visual models.

References

[1] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Süsstrunk, S. (2012). SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, *34*(11), 2274-2282. https://doi.org/10.1109/TPAMI.2012.120

[2] Arbelaez, P., Maire, M., Fowlkes, C., & Malik, J. (2010). Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, *33*(5), 898-916. https://doi.org/10.1109/TPAMI.2010.161

[3] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9650-9660). https://doi.org/10.1109/ICCV48922.2021.00951

[4] Chen, H., Jin, Y., Jin, G., Zhu, C., & Chen, E. (2021). Semisupervised semantic segmentation by improving prediction confidence. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(9), 4991-5003. https://doi.org/10.1109/TNNLS.2021.3066850

[5] Cho, J. H., Mall, U., Bala, K., & Hariharan, B. (2021). 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). https://doi.org/10.1109/CVPR46437.2021.01652

[6] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., ... & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3213-3223). https://doi.org/10.1109/CVPR.2016.350

[7] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, *88*(2), 303-338. https://doi.org/10.1007/s11263-009-0275-4

[8] Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *science*, *315*(5814), 972-976. https://doi.org/10.1126/science.1136800

[9] Ji, X., Henriques, J. F., & Vedaldi, A. (2019). Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9865-9874). https://doi.org/10.1109/ICCV.2019.00996

[10] Li, K., Wang, Z., Cheng, Z., Yu, R., Zhao, Y., Song, G., ... & Chen, J. (2023). Acseg: Adaptive conceptualization for unsupervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7162-7172). https://doi.org/10.1109/CVPR52729.2023.00692

[11] Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, *22*(8), 888-905. https://doi.org/10.1109/34.868688