SegUnify: A Unified Framework for Unsupervised Semantic Segmentation

Presenters: Prachi
Ekansh Juneja

OVERVIEW

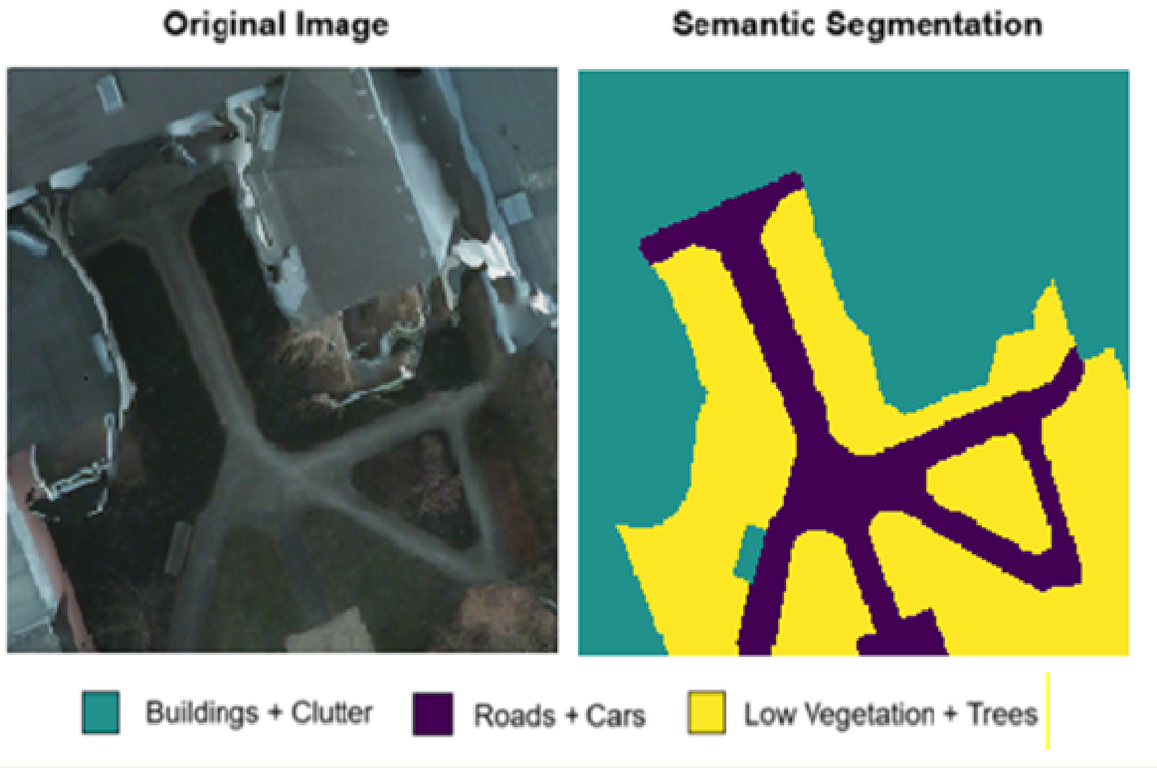
- Introduction
- Problem Statement
- Research Objectives
- Methodology
- Experiments Conducted
- Results Obtained
- Comparison with Existing Methods
- Conclusions and Future Scope
- References

INTRODUCTION

Semantic segmentation is dividing an image into meaningful regions and assigning a predefined category to each pixel.

Different applications of semantic segmentation include:

- Autonomous Driving
- Medical Imaging
- Satellite Imaging
- Robotics
- Security and Surveillance



PROBLEM STATEMENT

Semantic segmentation in supervised learning typically depends on large-scale labeled datasets, which are expensive and time-consuming to annotate. As described in a survey (Zhang et al., 2020), while supervised models can achieve high accuracy, the extensive requirement for pixel-level annotations severely limits scalability and practical deployment.

Unsupervised learning aims to reduce or eliminate the need for manual labeling. However, as highlighted by STEGO (Hamilton et al., 2022), current unsupervised segmentation methods often struggle with spatial consistency and handling complex scenes, resulting in suboptimal segmentation outputs. These challenges highlight the need for a robust, end-to-end framework that can handle diverse real-world images effectively and without external labeling.

RESEARCH OBJECTIVES

Unsupervised Framework

Eliminate the need for manual annotations in semantic segmentation.

Integrated Modules

Combine Feature Extraction, Dimensionality Reduction, and Labeling end-to-end for high-quality segmentation.

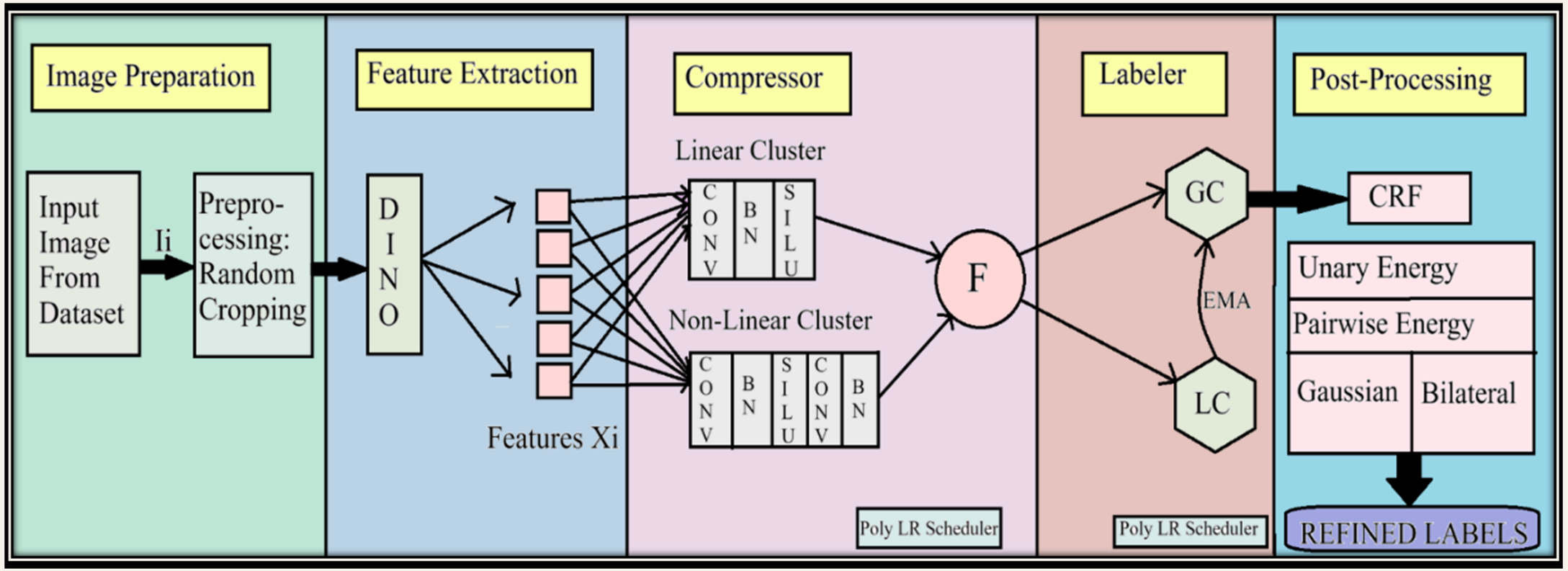
Scalability & Versatility

Maintain consistency across complex and diverse datasets, enabling robust performance in real-world applications.

Effective Real-World Adoption

Offers a scalable, unified approach adaptable to varying domains.

METHODOLOGY



METHODOLOGY

- The input images are given to the model. These images are preprocessed by being divided into five parts using the Five-Crop method. Then, the features are extracted using DINO, a pre-trained model on ImageNet. The parameters of DINO are kept frozen during the training process, and the extracted features are fed as input to the compressor module.

- The Compressor Module is used as a dimensionality reduction mechanism that maps high-dimensional extracted feature embeddings into a lower-dimensional space. Two main components are employed by the module: a linear cluster and an optional nonlinear cluster. Primitive features are extracted by the linear cluster, while more complex features are encapsulated by the nonlinear cluster.

- The linear cluster is defined by the equation :

$$F\_linear\ (X)=SiLU(BN(X * W\_conv\ ))$$

where X is the input feature map, Wconv represents the learnable weights of the convolutional layer, Sigmoid Linear Unit (SiLU) is the activation function and BN denotes batch normalization.

- The nonlinear cluster is defined by the equation :

$$F\_nonlinear\ (X)=SiLU(BN(SiLU(BN(X * W\_1\ )) * W\_2\ ))$$

where WI and W2 are the weights of two convolution layers in the sequential model respectively.

METHODOLOGY

- The combined output of the linear and nonlinear clusters is used to generate the final compressed feature representation F.
- The Labeler Module assigns labels to feature embeddings, F, by leveraging both local clusters (LC) and global clusters (GC). Local clusters are initialized using Xavier's normal initialization and dynamically updated during training, while global clusters are updated through an exponential moving average mechanism (EMA):

$$C\_global \leftarrow \alpha C\_global + (I-\alpha)C\_local$$

  where α is the momentum parameter.
- The local clusters try to align with the global cluster's learning. This alignment is driven by an energy minimization process, where the local cluster adapts to better fit the global cluster's output. This strategy ensures that semantic categories are learned to be predicted, even in the absence of ground truth.
- In the last step, the labels from the Labeler Module are enhanced using CRF to ensure refined segmentation outputs.

EXPERIMENTS CONDUCTED

- Model is evaluated on both original and augmented (horizontally flipped) images.
- Features are extracted from both versions of the images to capture diverse spatial cues.
- The Compressor Module reduces dimensionality while retaining spatial coherence.
- The Labeler Module assigns cluster-based labels to each pixel.
- A CRF (Conditional Random Field) refines boundaries and sharpens segmentation maps.

<u>Performance Metrics</u>

- Accuracy: Ratio of correctly labeled pixels over all pixels.
- mean Intersection over Union (mIoU): Average overlap between predicted segmentation and ground truth for each class.

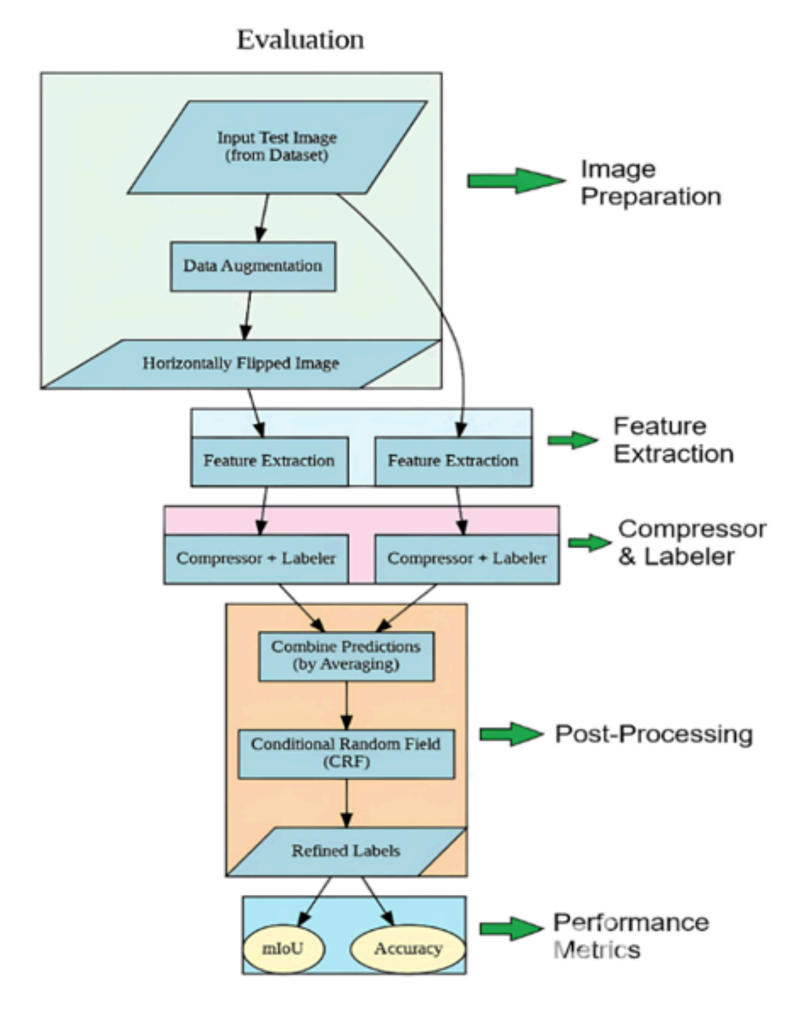
$$Accuracy = \frac{\sum_{i=1}^{C} TP(i)}{\sum_{i=1}^{C} (TP(i)+FP(i)+FN(i))}$$

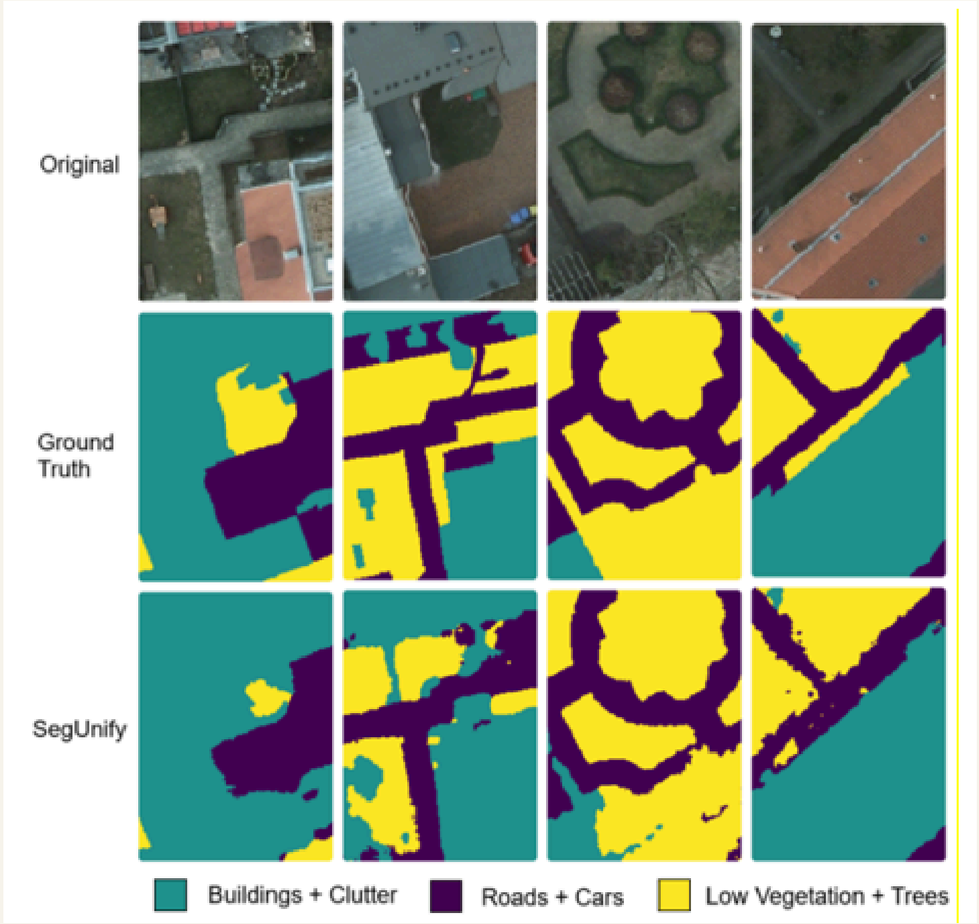
$$mIoU = \frac{1}{C}\sum_{i=1}^{C} \frac{TP(i)}{TP(i)+FP(i)+FN(i)}$$

<u>Dataset & Evaluation</u>

- Potsdam-3 Dataset: 8550 remote-sensing images divided into 3 major classes (Buildings & Clutter, Roads & Cars, Low Vegetation & Trees).
- Data Split: 7695 images for training, 855 for testing.
- Model tested on 855 images; each prediction refined via CRF and evaluated for Accuracy and mIoU.

EXPERIMENTS CONDUCTED



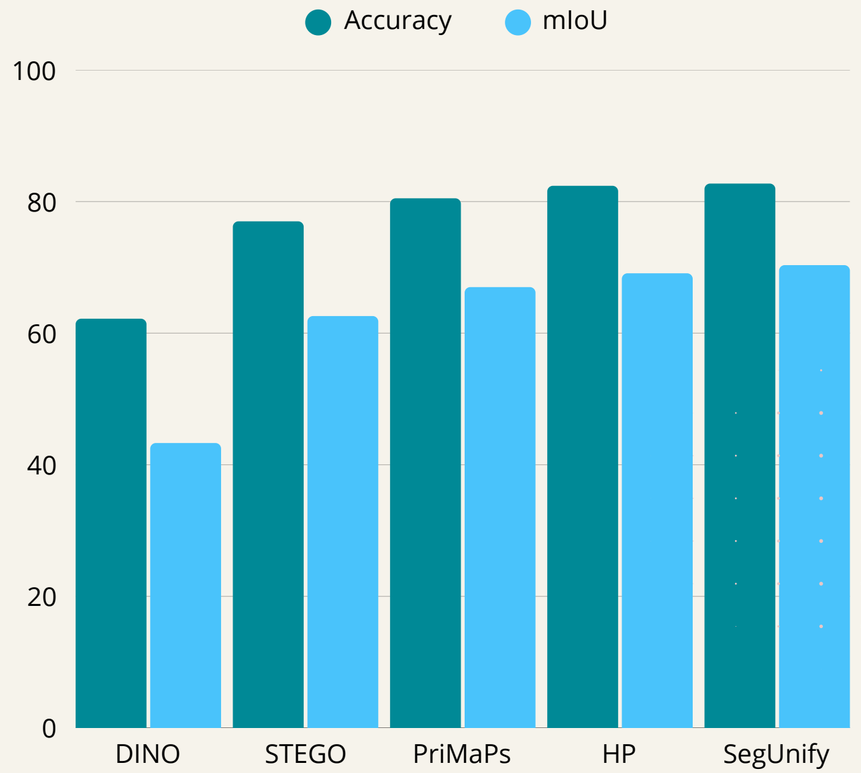


RESULTS OBTAINED

- **Proposed model SegUnify achieves 82.74% Accuracy and 70.34% mIoU on the Potsdam-3 dataset.**
- **Maintains higher coherence on complex boundaries compared to other methods.**

TABLE I. EXPERIMENTAL RESULTS ON POTSDAM-3 DATASET

Models	*Accuracy*	*mIoU*
DINO [11]	62.2	43.3
STEGO [12]	77.0	62.6
PriMaPs[32]	80.5	67.0
HP [33]	82.4	69.1
SegUnify (Proposed)	**82.74**	**70.34**



COMPARISON WITH EXISTING METHODS

S.No.	Title	Description	Outcome and Comparison
1.	CRF [7]	• Conditional Random Fields exploit local pixel similarities and spatial constraints for segmentation. • CRFs treat neighboring pixels with similar properties as belonging to the same class, promoting spatial consistency.	Outcome: Widely used for post-processing. Comparison: CRF helps refine boundaries but is not a full unsupervised pipeline. SegUnify replaces purely hand-crafted refinements with learned embeddings, yielding more robust semantic boundaries and higher accuracy.
2.	IIC [8]	• Invariant Information Clustering maximizes mutual information between different augmented views, facilitating unsupervised image classification and segmentation.	Outcome: Pioneered unsupervised clustering. Comparison: IIC ensures semantic coherence but lacks the multi-scale feature integration. SegUnify goes further by combining multi-level attention and region-based alignment, resulting in higher segmentation accuracy.
3.	PiCIE [9]	• Builds on IIC by enforcing photometric and geometric invariances at the pixel level, using stronger augmentation strategies.	Outcome: Achieved improved unsupervised semantic segmentation performance on datasets like PASCAL VOC and COCO-Stuff. Comparison: PiCIE is robust to transformations but does not fully leverage multi-scale transformer representations. SegUnify attains superior boundary precision and mIoU on Potsdam-3.

COMPARISON WITH EXISTING METHODS

S.No.	Title	Description	Outcome and Comparison
4.	VIT [10]	• Vision Transformers replace convolutional layers with self-attention, modeling long-range dependencies. Splits images into patches, enabling global context understanding.	Outcome: Provided strong feature representations for various vision tasks, but not specialized for unsupervised segmentation Comparison: SegUnify leverages multi-scale ViT features plus region-based refinement, achieving higher mIoU.
5.	DINO [11]	• Self-Distillation with No Labels uses a ViT backbone for unsupervised learning. Learns powerful feature representations by self-distillation, capturing high-level semantic features.	Outcome: Achieved 62.2% Accuracy, 43.3% mIoU on Potsdam-3. Often used as a feature extractor in unsupervised segmentation (e.g., STEGO) due to its robust semantic embeddings. Comparison: DINO excels in representation learning but lacks specialized segmentation modules. SegUnify significantly outperforms DINO (82.74% Acc, 70.34% mIoU) by integrating region-based attention and multi-scale consistency.
6.	STEGO [12]	• Unsupervised Semantic Segmentation by distilling feature correspondences. Leverages strong self-supervised features (DINO) and a contrastive objective to form discrete semantic clusters.	Outcome: Demonstrated good performance on COCO-Stuff and Cityscapes. Achieved 77.0% Accuracy, 62.6% mIoU on Potsdam-3. Comparison: STEGO improves boundary delineation over DINO, but SegUnify surpasses STEGO (82.74% Acc, 70.34% mIoU) with more refined region-level consistency and better handling across complex datasets.

CONCLUSIONS AND FUTURE SCOPE

Conclusion

- SegUnify effectively segments unlabeled images into meaningful semantic classes without external labels.
- Achieves efficient performance on Potsdam-3, surpassing popular unsupervised methods (STEGO, PriMaPs, HP, etc.).
- Experimental results show that Integration of Compressor Module, Labeler Module, and CRF refinement leads to highly coherent segmentation maps.

Future Scope

- Enhanced Feature Extraction
- Multiple Local Cluster Architecture
- Reducing Computational Complexity for Real time Inferences

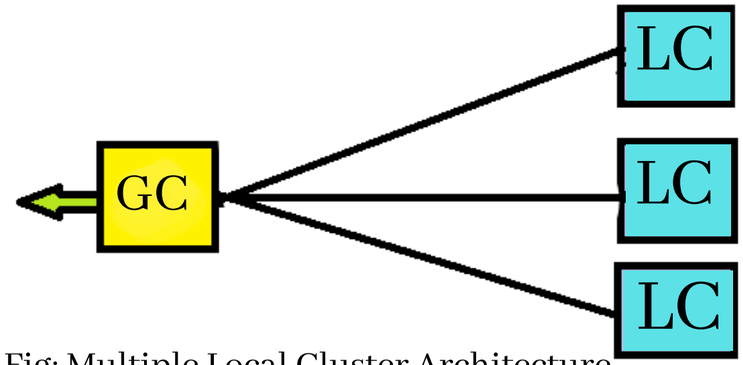


Fig: Multiple Local Cluster Architecture

REFERENCES

[1] A. Kulshreshtha and A. Nagpal, "Brain image segmentation using variation in structural elements of morphological operators," International Journal of Information Technology, vol. 15, no. 4, pp. 2283–2291, 2023.
[2] M. L. Silvoster, R. Mathusoothana, and S. Kumar, "Watershed based algorithms for the segmentation of spine MRI," International Journal of Information Technology, vol. 14, no. 3, pp. 1343–1353, 2022.
[3] P. Sahare, J. V. Tembhurne, M. R. Parate, T. Diwan, and S. B. Dhok, "Script independent text segmentation of document images using graph network based shortest path scheme," International Journal of Information Technology, vol. 15, no. 4, pp. 2247–2261, 2023.
[4] S. Vignesh, M. Savithadevi, M. Sridevi, and R. Sridhar, "A novel facial emotion recognition model using segmentation VGG-19 architecture," International Journal of Information Technology, vol. 15, no. 4, pp. 1777–1787, 2023.
[5] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," Advances in neural information processing systems, vol. 34, pp. 12077–12090, 2021.
[6] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 6881–6890.
[7] J. Lafferty, A. McCallum, F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in ICML, Williamstown, MA, 2001, p. 3.
[8] X. Ji, J. F. Henriques, and A. Vedaldi, "Invariant information clustering for unsupervised image classification and segmentation," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 9865–9874.
[9] J. H. Cho, U. Mall, K. Bala, and B. Hariharan, "PiCIE: Unsupervised semantic segmentation using invariance and equivariance in clustering," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16794–16804.
[10] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.
[11] M. Caron et al., "Emerging properties in self-supervised vision transformers," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 9650–9660.
[12] M. Hamilton, Z. Zhang, B. Hariharan, N. Snavely, and W. T. Freeman, "Unsupervised semantic segmentation by distilling feature correspondences," arXiv preprint arXiv:2203.08414, 2022.
[13] S. Beucher and F. Meyer, "Segmentation: The Watershed Transformation. Mathematical Morphology in Image Processing," Optical Engineering, vol. 34, pp. 433–481, Jan. 1993.
[14] M. Mancas, B. Gosselin, and B. Macq, "Segmentation using a region-growing thresholding," presented at the Electronic Imaging 2005, E. R. Dougherty, J. T. Astola, and K. O. Egiazarian, Eds., San Jose, CA, Mar. 2005, p. 388. doi: 10.1117/12.587995.
[15] F. U. Siddiqui and A. Yahya, Clustering Techniques for Image Segmentation. Cham: Springer International Publishing, 2022. doi: 10.1007/978-3-030-81230-0.
[16] T. V. Le, C. A. Kulikowski, and I. B. Muchnik, "A Graph-Based Approach for Image Segmentation," in Advances in Visual Computing, vol. 5358, in Lecture Notes in Computer Science, vol. 5358. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 278–287. doi: 10.1007/978-3-540-89639-5_27.
[17] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-Based Convolutional Neural Network for Whole Slide Tissue Image Classification," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2424–2433. doi: 10.1109/CVPR.2016.266.

REFERENCES

[18] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," Advances in neural information processing systems, vol. 24, 2011.
[19] A. Arnab et al., "Conditional Random Fields Meet Deep Neural Networks for Semantic Segmentation: Combining Probabilistic Graphical Models with Deep Learning for Structured Prediction," IEEE Signal Processing Magazine, vol. 35, no. 1, pp. 37–52, 2018, doi: 10.1109/MSP.2017.2762355.
[20] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," 2014, arXiv. doi: 10.48550/ARXIV.1411.4038.
[21] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," Oct. 10, 2016, arXiv: arXiv:1511.00561. doi: 10.48550/arXiv.1511.00561.
[22] R. Harb and P. Knöbelreiter, "Infoseg: Unsupervised semantic image segmentation with mutual information maximization," in DAGM German Conference on Pattern Recognition, Springer, 2021, pp. 18–32.
[23] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, and L. Van Gool, "Unsupervised semantic segmentation by contrasting object mask proposals," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10052–10062.
[24] M. Kairanbay, J. See, and L.-K. Wong, "Aesthetic Evaluation of Facial Portraits Using Compositional Augmentation for Deep CNNs," in Computer Vision – ACCV 2016 Workshops, vol. 10117, C.-S. Chen, J. Lu, and K.-K. Ma, Eds., in Lecture Notes in Computer Science, vol. 10117. , Cham: Springer International Publishing, 2017, pp. 462–474. doi: 10.1007/978-3-319-54427-4_34.
[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Commun. ACM, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.
[26] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in Proceedings of the thirteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.
[27] M. B. Perry, "The Exponentially Weighted Moving Average," in Wiley Encyclopedia of Operations Research and Management Science, 1st ed., Wiley, 2011. doi: 10.1002/9780470400531.eorms0314.
[28] M. Nawaz et al., "Unravelling the complexity of Optical Coherence Tomography image segmentation using machine and deep learning techniques: A review," Computerized Medical Imaging and Graphics, p. 102269, 2023.
[29] Z. Wang et al., "Revisiting evaluation metrics for semantic segmentation: Optimization and evaluation of fine-grained intersection over union," Advances in Neural Information Processing Systems, vol. 36, 2024.
[30] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," Advances in neural information processing systems, vol. 32, 2019.
[31] P. Mishra and K. Sarawadekar, "Polynomial Learning Rate Policy with Warm Restart for Deep Neural Network," in TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), 2019, pp. 2087–2092. doi: 10.1109/TENCON.2019.8929465.
[32] O. Hahn, N. Araslanov, S. Schaub-Meyer, and S. Roth, "Boosting unsupervised semantic segmentation with principal mask proposals," arXiv preprint arXiv:2404.16818, 2024.
[33] H. S. Seong, W. Moon, S. Lee, and J.-P. Heo, "Leveraging hidden positives for unsupervised semantic segmentation," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 19540–19549.

SegUnify| 2025

THANK YOU

Presented By: Prachi
Ekansh Juneja