Multi-Scale Unsupervised Image Segmentation with Inception-Based Feature Clustering

Asif Azad
asifazad0178@gmail.com
Bangladesh University of Engineering and Technology

Wasif Hamid
wsf.hmd99@gmail.com
Bangladesh University of Engineering and Technology

ABSTRACT

This study presents an unsupervised image segmentation framework leveraging an InceptionNet-inspired architecture to cluster image pixels without labeled data. By integrating multi-scale feature extraction with differentiable clustering, the proposed method addresses the challenges of feature representation, spatial coherence, and adaptability in diverse datasets. Evaluated on PASCAL VOC 2012 and COCO-Stuff datasets, the model demonstrates robust segmentation performance, overcoming limitations of traditional methods reliant on hand-crafted features or fixed boundaries. The framework's end-to-end optimization and clustering approach mark a significant step toward scalable and efficient unsupervised segmentation. Here is the project repository: GitHub Repository.

KEYWORDS

Image Segmentation, Convolutional neural networks, Unsupervised learning, Computer Vision, Feature clustering

ACM Reference Format:
Asif Azad and Wasif Hamid. 2025. Multi-Scale Unsupervised Image Segmentation with Inception-Based Feature Clustering. In . ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Image segmentation is a critical task in computer vision, where the goal is to partition an image into distinct regions, each corresponding to meaningful objects or areas of interest. This process has a wide array of applications, including medical imaging, autonomous vehicles, and object recognition, making it a fundamental area of research.

While supervised segmentation methods, such as those based on convolutional neural networks (CNNs), have achieved remarkable success, they often rely on large annotated datasets. The collection and labeling of such datasets is time-consuming, expensive, and prone to human error. Moreover, supervised semantic segmentation methods have limitations in handling unknown object classes since they are constrained to classify each pixel into one of the predefined categories.As a result, there is growing interest in unsupervised approaches to image segmentation, which aim to cluster pixels or regions into segments without requiring labeled data.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

The challenge in unsupervised segmentation lies in accurately representing the complex features of an image, such as textures, colors, and spatial relationships, in a way that supports effective clustering of pixels. Traditional methods often rely on fixed heuristics or hand-crafted features, which may lack the flexibility to adapt to diverse image content. On the other hand, deep learning techniques, particularly convolutional neural networks (CNNs), have demonstrated remarkable success in learning rich hierarchical representations of image data. Despite their widespread use in supervised tasks, the potential of CNNs for unsupervised segmentation remains underexplored.

We take a closer look at what makes good cluster labels for effective image segmentation. Similar to earlier research on unsupervised image segmentation [16], [19], we assume that the best segmentation closely resembles what a human would naturally do. If you ask someone to segment an image, they'd likely create regions that represent whole objects or significant parts of them. Objects often have large areas with similar colors or consistent texture patterns, so it makes sense to group nearby pixels with similar colors or textures into the same cluster.

At the same time, to separate different objects, it is important to assign different cluster labels to neighboring pixels that look noticeably different. This approach helps create clear boundaries between objects, matching how humans naturally perceive and divide visual scenes.

2 LITERATURE REVIEW

Supervised learning techniques for image segmentation have seen significant progress with convolutional neural networks (CNNs) offering state-of-the-art results. Methods such as Fully Convolutional Networks (FCN) and DeepLab architectures have pushed boundaries but rely heavily on annotated datasets. Weakly-supervised approaches, using image-level labels or bounding boxes, mitigate annotation costs but still depend on partial supervision.

In contrast, unsupervised segmentation methods utilize intrinsic image features such as color, texture, and spatial coherence to group pixels. Traditional approaches, like the graph-based segmentation (GS)[5] algorithm and mean-shift clustering[13], emphasize spatial relationships but often lack flexibility when applied to diverse image content.

Recent advancements, such as Deep Embedded Clustering (DEC) [18] and invariant information clustering, leverage deep learning to jointly optimize feature representations and cluster assignments. MsLRR[11], a multi-scale learning algorithm, combines local feature extraction with superpixel-based refinement. However, methods

that rely on superpixels impose fixed boundaries, limiting adaptability. Other approaches, like W-Net[17], aim for end-to-end unsupervised learning by estimating segmentation and restoring input images.

The work of Kanezaki[8] contributed significantly to unsupervised segmentation by proposing a CNN-based approach that uses superpixels to enforce spatial constraints. This method jointly optimized pixel cluster labels and network parameters, showing that a backpropagation-based training framework can yield meaningful segmentations. However, its reliance on fixed superpixels limited the adaptability of segment boundaries.

Building on this, Kim et al.[9] introduced an end-to-end framework that replaced superpixel dependency with a spatial continuity loss. This innovation allowed segment boundaries to adapt dynamically during training, significantly improving flexibility and segmentation accuracy. Moreover, their approach extended to include additional applications such as user-guided scribble-based segmentation and unseen image segmentation using pre-trained weights, demonstrating the versatility of the method.

More recent work combined reconstruction losses with additional clustering modules[20] or improved superpixel segmentation integration[10]. Despite these advancements, many methods continue to face challenges like over-segmentation and inconsistency in features within objects, fixed segment boundaries, suboptimal cluster initialization, and limited integration of spatial and feature-based constraints. Addressing these gaps is critical for robust, generalizable unsupervised segmentation, paving the way for more adaptive and efficient models.

3 METHODOLOGY

3.1 Proposed Architecture

The proposed method incorporates an advanced InceptionNet-based architecture to achieve superior unsupervised segmentation performance. The InceptionNet design enhances feature extraction by utilizing multi-scale convolutional operations within the same layer. Key components of the architecture include:

3.1.1 Feature Extraction Module. The initial stage uses a standard convolutional block to preprocess the input image. Each block consists of a convolutional layer followed by batch normalization and ReLU activation to ensure effective feature extraction and normalization.

3.1.2 Inception Block. Inspired by the InceptionNet design, the architecture employs multiple branches within a single block to process features at different scales:

- **Branch 1:** A 1x1 convolutional layer captures fine-grained features.
- **Branch 2:** A reduction 1x1 convolutional layer is followed by a 3x3 convolution for medium-scale feature extraction.
- **Branch 3:** Similar to Branch 2, but uses a 5x5 convolution to extract broader contextual features.
- **Branch 4:** A max-pooling operation, followed by a 1x1 convolution, ensures that spatial information is preserved.

The outputs of all branches are concatenated along the channel dimension, allowing the network to simultaneously capture multi-scale features.

3.1.3 Intermediate Layers. To further refine the extracted features, successive Inception blocks are stacked, each followed by batch normalization. This design enables deeper feature hierarchies while mitigating the vanishing gradient problem.

3.1.4 Clustering Module. The final layers project the extracted features into a high-dimensional space, where clustering is performed. A 1x1 convolutional layer reduces the feature dimensionality, followed by batch normalization to prepare for clustering.

3.1.5 Differentiable Clustering. The clustering operation assigns pixel-level cluster IDs based on feature similarity. An argmax function is applied to normalize the clustering assignments, facilitating end-to-end optimization.

By leveraging the InceptionNet architecture, the proposed model effectively captures multi-scale features and ensures robust segmentation across diverse image datasets. This architecture is particularly suited for unsupervised segmentation due to its ability to adaptively focus on different feature scales.

3.2 Dataset

The experiments utilized the PASCAL VOC 2012 and COCO-Stuff datasets, which are some of the biggest and most well regarded benchmarks for image segmentation tasks.

3.2.1 PASCAL VOC 2012.

- **Description:** The PASCAL VOC dataset contains 11,530 images annotated with 20 object categories and a background class. It is widely used for object detection, classification, and segmentation tasks. For segmentation, it includes pixel-wise annotations that facilitate fine-grained analysis.
- **Statistics:** The dataset consists of 1,464 images for training, 1,449 for validation, and 1,456 for testing. The images are diverse, spanning indoor and outdoor scenes with complex backgrounds, making it challenging for segmentation models.

3.2.2 COCO-Stuff.

- **Description:** The COCO-Stuff dataset extends the popular COCO dataset by adding dense pixel-level annotations for "stuff" classes such as grass, sky, and water, along with the original object ("thing") classes. It contains 164,000 images with 91 "thing" categories and 91 "stuff" categories, resulting in a total of 182 classes.
- **Statistics:** The dataset includes 118,000 training images, 5,000 validation images, and 41,000 test images. Its large size and variety make it a challenging and comprehensive benchmark for segmentation tasks.

3.2.3 Dataset Analysis.

- **PASCAL VOC 2012:** This dataset provides a compact yet diverse dataset ideal for prototyping and testing segmentation algorithms. The relatively small size allows for rapid experimentation, but its complexity ensures that only robust methods perform well.
- **COCO-Stuff:** With its significantly larger size and inclusion of "stuff" categories, this dataset challenges models to handle dense, cluttered scenes and subtle boundaries between

"things" and "stuff." It is particularly suitable for testing the scalability and generalization of segmentation models.

By combining insights from these datasets, our experiments evaluated the proposed model's ability to handle diverse scenarios, ranging from simple object segmentation to complex scenes with mixed categories.

3.3 Training Details

As an unsupervised model, the training process is unique in that it requires learning on each image separately. The training was conducted using a PyTorch-based implementation on an NVIDIA GeForce RTX 4070 GPU with CUDA 12.4, ensuring efficient computation. The key details of the training process are as follows:

(1) **Initialization**: The model weights were initialized using Xavier initialization to ensure efficient gradient flow. The network utilized an Inception-based architecture with the number of convolutional layers (`-nConv`) set to 2.
(2) **Parameter Configuration**: The following parameters were used during training:
   - **Number of Channels (`-nChannel`)**: Set to 100 to define the dimensionality of feature representation.
   - **Learning Rate (`-lr`)**: Configured at 0.1 to balance convergence speed and stability.
   - **Maximum Iterations (`-maxIter`)**: Adjusted to 150 for validation runs, while the default was set to 1000 for typical training sessions.
   - **Minimum Labels (`-minLabels`)**: Configured to a minimum of 3 labels to ensure segmentation granularity.
   - **Step Sizes for Loss Components (`-stepsize_sim` and `-stepsize_con`)**: Both were set to 1, assigning equal importance to feature similarity and spatial continuity losses.
(3) **Label Prediction**: For each input image, pixel labels were initialized randomly and iteratively updated using the argmax operation applied to feature responses.
(4) **Loss Optimization**: The network optimized a composite loss function comprising:
   - **Feature Similarity Loss**: Ensures that pixels within the same cluster have similar features.
   - **Spatial Continuity Loss**: Encourages spatially adjacent pixels to belong to the same cluster, ensuring smooth and coherent segmentation boundaries.

Each image was processed independently, enabling the model to adapt to diverse input scenarios. The unsupervised nature of the training requires significant computational resources for segmentation tasks, but the modular approach ensures flexibility and scalability.

3.4 Evaluation Metrics

The performance of the segmentation model was evaluated using four metrics:

(1) **Mean Intersection over Union (mIoU)**: Measures the overlap between predicted and ground truth segments, averaged across all classes. A higher mIoU indicates better alignment with ground truth.
(2) **Accuracy**: Calculates the percentage of correctly labeled pixels. While straightforward, this metric may not fully capture segmentation quality in complex scenes.
(3) **Homogeneity Score**: Evaluates whether each cluster contains only pixels belonging to a single class, emphasizing the purity of clusters.
(4) **Normalized Mutual Information (NMI)**: Measures the amount of shared information between the predicted segmentation and the ground truth. NMI ranges from 0 to 1, where 1 indicates perfect agreement between the predicted and ground truth segmentations. It is particularly useful for evaluating clustering algorithms in an unsupervised context, as it balances the impact of cluster size and label matching.

4 EXPERIMENTAL RESULTS

5 RESULTS

This section presents a comprehensive quantitative evaluation of our proposed unsupervised segmentation approach. The performance is measured using four key metrics: mean Intersection over Union (mIoU), Accuracy, Normalized Mutual Information (NMI), and Homogeneity Score. The results are analyzed across two architectures—InceptionNet and Vanilla CNN—on two widely-used benchmarks: PASCAL VOC 2012 and COCO-Stuff 2017 datasets. A detailed comparison with existing state-of-the-art methods is also included.

5.1 Performance on PASCAL VOC 2012

The performance of the models on the PASCAL VOC 2012 validation set, consisting of 1449 images, is shown in Table 1. Vanilla CNN achieved an mIoU of 0.2585, surpassing InceptionNet, which recorded 0.2116. However, InceptionNet exhibited higher accuracy (0.7153) compared to Vanilla CNN (0.6539). This suggests that Vanilla CNN's simpler architecture may favor pixel-level segmentation (reflected in mIoU), while InceptionNet's multi-scale feature extraction offers superior accuracy by capturing more robust global features. Additionally, InceptionNet demonstrated a higher Homogeneity Score, emphasizing its strength in clustering features more cohesively, particularly in scenarios involving complex label distributions.

Table 1: Performance metrics (mIoU, Accuracy, NMI, and Homogeneity Score) of InceptionNet and Vanilla CNN on the validation set (1449 images) of the PASCAL VOC 2012 dataset.

Architecture	mIoU	Accuracy	NMI	Homogeneity Score
Vanilla CNN	0.2585	0.6539	0.1989	0.3745
InceptionNet	0.2116	0.7153	0.1959	0.4836

The NMI scores for both architectures remained close, indicating comparable performance in identifying clusters aligned with ground truth. However, InceptionNet's higher Homogeneity Score suggests that its features are more consistently grouped into meaningful clusters, which is a valuable attribute for segmentation tasks.

5.2 Performance on COCO-Stuff 2017

For the COCO-Stuff 2017 dataset, comprising 5000 images, Table 2 illustrates the comparative performance of the two architectures. InceptionNet achieved a higher mIoU (0.2233) than Vanilla CNN (0.2113), reflecting its robustness in handling the increased complexity of this dataset. Additionally, InceptionNet excelled in Accuracy, NMI, and Homogeneity Score, with improvements of 4.1%, 3.4%, and 5.9%, respectively, over Vanilla CNN.

Table 2: Performance metrics (mIoU, Accuracy, NMI, and Homogeneity Score) of InceptionNet and Vanilla CNN on the validation set (5000 images) of the COCO-Stuff 2017 dataset.

Architecture	mIoU	Accuracy	NMI	Homogeneity Score
Vanilla CNN	0.2113	0.3781	0.3545	0.5015
InceptionNet	0.2233	0.4188	0.3666	0.5602

These results demonstrate InceptionNet's ability to extract multi-scale features, which are crucial for accurate segmentation in datasets with diverse and complex scenes. The higher NMI and Homogeneity Score indicate improved clustering and segmentation consistency, making it particularly well-suited for COCO-Stuff 2017.

5.3 Comparison with State-of-the-Art Methods

We compare the performance of our InceptionNet architecture against existing state-of-the-art methods on the COCO-Stuff 2017 validation dataset. Table 3 provides a quantitative overview, reporting Accuracy and mIoU for each method.

ResNet50 and MoCoV2 demonstrate relatively low mIoU values (8.9 and 10.4, respectively), reflecting their limited capacity for segmentation in unsupervised scenarios. Traditional feature-based methods, such as SIFT, also underperform due to their inability to leverage feature hierarchies learned during training. Modern approaches like DINO and AC improve segmentation performance, with DINO achieving an mIoU of 9.6 and AC reaching an Accuracy of 30.8%.

Advanced clustering and self-supervised methods such as InMARS, IIC, and MDC achieve further improvements, with MDC surpassing earlier techniques with a notable Accuracy of 32.2% and mIoU of 9.8. PiCIE, leveraging a more robust clustering strategy, achieves an impressive mIoU of 13.8 and Accuracy of 41.1%, making it one of the strongest methods among the prior works.

Our proposed InceptionNet architecture outperforms all previous methods, achieving the highest Accuracy of **41.88%** and mIoU of **22.33%**. These results highlight the superiority of our approach in handling complex segmentation tasks, particularly in datasets with intricate label distributions and diverse visual contexts.

6 DISCUSSION

The findings of this study highlight the significance of architecture design and dataset characteristics in unsupervised segmentation. InceptionNet's superior performance on the COCO-Stuff 2017 dataset demonstrates the effectiveness of multi-scale feature extraction, which allows the model to capture complex spatial relationships and feature hierarchies. This is particularly advantageous for datasets with diverse scenes and intricate label distributions. The higher

Table 3: Quantitative results on the COCO-Stuff validation dataset (Caesar et al., 2018).

Methods	Accuracy (%)	mIoU (%)
ResNet50 [6]	24.6	8.9
MoCoV2 [3]	25.2	10.4
DINO [2]	30.5	9.6
Deep Cluster [1]	19.9	–
SIFT [12]	20.2	–
AC [15]	30.8	–
InMARS [14]	31.0	–
IIC [7]	21.8	6.7
MDC [4]	32.2	9.8
PiCIE [4]	41.1	13.8
InceptionNet (Ours)	**41.88**	**22.33**

NMI and Homogeneity Scores achieved by InceptionNet reinforce its capability to cluster features cohesively, which is critical for segmentation tasks.

On the other hand, Vanilla CNN's higher mIoU on the PASCAL VOC 2012 dataset suggests that simpler architectures may sometimes outperform more complex ones in datasets with fewer images and simpler scenes. This indicates a trade-off between capturing global features and excelling in pixel-level segmentation. The comparison with state-of-the-art methods further highlights the competitiveness of our approach, particularly in terms of accuracy and mIoU, where it outperformed existing models like PiCIE-H by a significant margin.

These results suggest that the choice of architecture should be guided by the dataset's complexity and application requirements. The demonstrated improvements in clustering and segmentation accuracy pave the way for future work exploring multi-scale architectures in unsupervised learning scenarios.

7 CONCLUSION

The proposed InceptionNet-based architecture provides a novel and efficient approach for unsupervised image segmentation by capturing multi-scale features and employing adaptive clustering techniques. Comprehensive evaluations on diverse datasets validate the model's ability to generalize across complex scenarios. This study bridges a critical gap in the field by demonstrating the potential of deep convolutional networks in unsupervised segmentation, paving the way for future research into more flexible and scalable models. Despite its success, areas like initialization strategies and cluster refinement warrant further exploration.

REFERENCES

[1] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*. 132–149.

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 9630–9640. https://doi.org/10.1109/ICCV48922.2021.00951

[3] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297* (2020).

[4] Jang Hyun Cho, Utkarsh Mall, Kavita Bala, and Bharath Hariharan. 2021. Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*

Recognition. 16794–16804.

[5] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. 2004. Efficient Graph-Based Image Segmentation. *International Journal of Computer Vision* 59 (2004), 167–181. https://api.semanticscholar.org/CorpusID:207663697

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[7] Xu Ji, Joao F Henriques, and Andrea Vedaldi. 2019. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9865–9874.

[8] Asako Kanezaki. 2018. Unsupervised image segmentation by backpropagation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1543–1547.

[9] Wonjik Kim, Asako Kanezaki, and Masayuki Tanaka. 2020. Unsupervised learning of image segmentation based on differentiable feature clustering. *IEEE Transactions on Image Processing* 29 (2020), 8055–8068.

[10] Qinghong Lin, Weichan Zhong, and Jianglin Lu. 2021. Deep Superpixel Cut for Unsupervised Image Segmentation. *2020 25th International Conference on Pattern Recognition (ICPR)* (2021), 8870–8876. https://api.semanticscholar.org/CorpusID:232170352

[11] Xiaobai Liu, Qian Xu, Jiayi Ma, Hai Jin, and Yanduo Zhang. 2014. MsLRR: A Unified Multiscale Low-Rank Representation for Image Segmentation. *IEEE Transactions on Image Processing* 23 (2014), 2159–2167. https://api.semanticscholar.org/CorpusID:1409200

[12] David G Lowe. 1999. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, Vol. 2. Ieee, 1150–1157.

[13] J. MacQueen. 1967. Some methods for classification and analysis of multivariate observations. https://api.semanticscholar.org/CorpusID:6278891

[14] S Ehsan Mirsadeghi, Ali Royat, and Hamid Rezatofighi. 2021. Unsupervised image segmentation by mutual information maximization and adversarial regularization. *IEEE Robotics and Automation Letters* 6, 4 (2021), 6931–6938.

[15] Yassine Ouali, Céline Hudelot, and Myriam Tami. 2020. Autoregressive unsupervised image segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*. Springer, 142–158.

[16] Ranjith Unnikrishnan, Caroline Pantofaru, and Martial Hebert. 2007. Toward Objective Evaluation of Image Segmentation Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 6 (2007), 929–944. https://doi.org/10.1109/TPAMI.2007.1046

[17] Xide Xia and Brian Kulis. 2017. W-Net: A Deep Model for Fully Unsupervised Image Segmentation. *ArXiv* abs/1711.08506 (2017). https://api.semanticscholar.org/CorpusID:12565210

[18] Junyuan Xie, Ross B. Girshick, and Ali Farhadi. 2015. Unsupervised Deep Embedding for Clustering Analysis. *ArXiv* abs/1511.06335 (2015). https://api.semanticscholar.org/CorpusID:6779105

[19] Allen Y. Yang, John Wright, Yi Ma, and S. Shankar Sastry. 2008. Unsupervised segmentation of natural images via lossy data compression. *Computer Vision and Image Understanding* 110, 2 (2008), 212–225. https://doi.org/10.1016/j.cviu.2007.07.005

[20] Lei Zhou and Weiyufeng Wei. 2020. DIC: Deep Image Clustering for Unsupervised Image Segmentation. *IEEE Access* 8 (2020), 34481–34491. https://api.semanticscholar.org/CorpusID:211690125