GroupViT: Text-Based Training for Visual Segmentation Without Pixel-Level Labels

Group 11
CVPR Submission
Jio Institute, Sector 4, Ulwe, Navi Mumbai

Abstract

This paper presents GroupViT (Grouping Vision Transformer), a novel approach to semantic segmentation that utilizes text supervision instead of pixel-level annotations. This method addresses the challenges of traditional segmentation, such as the high cost of detailed labeling and limited generalization to unseen categories. GroupViT introduces a hierarchical grouping mechanism within a Vision Transformer (ViT) framework, leveraging large-scale image-text datasets to identify meaningful segments without manual labels. Training relies on contrastive learning, optimizing the model's ability to differentiate correct and incorrect image-text pairs. Our model achieves competitive performance on established benchmarks like PASCAL VOC 2012 (52.3% mIoU) and PASCAL Context (22.4% mIoU), highlighting its ability to generalize beyond the limitations of traditional annotation-heavy methods.

1. Introduction

Semantic segmentation is the task of partitioning an image into semantically meaningful regions, requiring high granularity, often at the pixel level. It is a crucial component in a wide range of computer vision applications, such as autonomous driving, medical imaging, and object detection. However, traditional deep learning approaches for semantic segmentation face two significant challenges:

1.1. Challenges in Traditional Segmentation

High Cost of Annotations: Creating pixel-level labels for large datasets requires substantial time and effort from human annotators, resulting in high costs. This limits the scalability of traditional supervised learning approaches, especially when dealing with diverse and complex datasets.

Restricted Generalization: Models trained using annotated data often lack the ability to generalize to unseen categories, necessitating retraining or fine-tuning for new tasks. This reduces the flexibility and adaptability of conventional segmentation models, making them unsuitable for dynamic environments.

1.2. Text Supervision: A New Paradigm

Recent advancements in text-supervised visual learning offer an alternative to pixel-level annotation. By using descriptive captions associated with images, models can capture a broader understanding of content without requiring precise annotations.

This study introduces GroupViT, a novel model that combines the flexibility of Vision Transformers with a hierarchical grouping mechanism. By training on large-scale image-text pairs, GroupViT enables zero-shot semantic segmentation, allowing the model to categorize previously unseen objects based on textual descriptions. This approach promises to overcome the challenges faced by traditional methods and paves the way for more flexible segmentation solutions.

2. Related Work

2.1. Supervised Approaches

Semantic segmentation has traditionally relied on fully supervised approaches, which require a vast amount of labeled data for accurate predictions.

Fully Convolutional Networks (FCNs) introduced an end-to-end methodology for pixel-wise predictions, leading to a significant increase in segmentation accuracy. These models, however, necessitate extensive pixel-level annotations, limiting scalability.

DeepLab and its variants refined segmentation accuracy by introducing concepts like atrous convolution and multi-scale context aggregation. Despite achieving state-of-the-art performance, the dependence on labeled data remains a bottleneck.

2.2. Weakly Supervised and Text-Supervised Learning

Weakly supervised learning has gained attention for reducing annotation costs by using weaker labels, such as

image-level tags or bounding boxes. However, this often results in lower accuracy.

Text-Supervised Learning has emerged as a promising direction, with models like CLIP demonstrating the effectiveness of training on image-text pairs. This approach enables zero-shot transfer to unseen categories, laying the foundation for GroupViT's development in the segmentation domain.

2.3. Vision Transformers in Segmentation

Vision Transformers (ViTs) have recently transformed the landscape of visual learning due to their ability to capture complex patterns through self-attention mechanisms.

ViTs for Segmentation have shown remarkable adaptability in learning both local and global features, enabling more sophisticated understanding of image content. GroupViT leverages ViTs to facilitate text-driven segmentation by employing hierarchical grouping, allowing the model to accurately identify segments without reliance on grids.

3. Method

GroupViT integrates the strengths of Vision Transformers and hierarchical grouping to perform semantic segmentation with minimal supervision. Here's a detailed breakdown of the model's architecture:

3.1. Vision Transformer (ViT) Base

The backbone of GroupViT is a Vision Transformer (ViT) that divides input images into patches, typically 16x16 pixels, treating each patch as an individual token. These tokens are passed through a series of Transformer layers, allowing the model to learn relationships between patches.

Self-Attention Mechanism: ViTs utilize self-attention to identify dependencies among patches, enabling the model to focus on relevant features while ignoring irrelevant details. This flexibility allows the model to understand both fine-grained details and the broader context of an image.

3.2. Hierarchical Grouping

GroupViT employs a hierarchical grouping strategy to merge visual tokens progressively, facilitating accurate segmentation without pixel-level labels.

Multi-Stage Grouping: In the early stages, the model targets smaller, detailed regions (e.g., facial features), and merges them into larger, more coherent segments in subsequent stages. This multi-scale approach captures object details at different resolutions, leading to more accurate segmentation.

Adaptive Grouping: GroupViT's grouping mechanism allows segments to form organically, adjusting to the content of the image rather than adhering to a fixed grid. This leads to improved segmentation quality, particularly in complex and cluttered scenes.

3.3. Grouping Block

The Grouping Block is a core innovation in GroupViT, facilitating the dynamic merging of visual tokens based on learned features.

Gumbel-Softmax Mechanism: To make the grouping process differentiable, GroupViT utilizes Gumbel-Softmax, enabling backpropagation during training. This mechanism allows the model to assign visual tokens to specific groups dynamically, refining segments with each iteration.

Iterative Refinement: GroupViT iteratively refines segment boundaries through multiple grouping stages. This process enhances the model's ability to handle varied and complex visual content, producing more accurate and semantically meaningful segments.

3.4. Text Supervision via Contrastive Learning

GroupViT leverages contrastive learning with paired image-text data to build strong associations between visual features and textual descriptions. This is done by training the model to distinguish between correct and incorrect image-text pairs.

- **Positive Pairs**: These consist of image-text pairs that accurately describe the image content, bringing their feature representations closer together.
- **Negative Pairs**: These are incorrect pairs that do not match, encouraging the model to push apart their feature representations.

By learning these associations, GroupViT can perform zero-shot segmentation, categorizing unseen objects based solely on their textual descriptions.

4. Experiments

To validate the effectiveness of GroupViT, we conducted extensive experiments on standard benchmarks, measuring segmentation performance under different conditions.

4.1. Datasets

PASCAL VOC 2012: This dataset includes 20 object categories plus a background class, providing a variety of everyday objects for testing segmentation accuracy.

PASCAL Context: A more challenging dataset with 59 object categories, including detailed background elements. It is designed to assess the model's ability to handle complex and densely labeled scenes.

Table 1. GroupViT Performance on PASCAL VOC 2012

Model	Training Method	mIoU (%)
FCN-8s (Baseline)	Supervised	62.7
DeepLabV3+	Supervised	79.3
GroupViT (Ours)	Text Supervision	52.3

4.2. Evaluation Metrics

We evaluated the performance of GroupViT using the following metrics:

Mean Intersection over Union (mIoU): This metric is used to assess segmentation quality by measuring the overlap between predicted segments and ground truth. A higher mIoU indicates a better alignment of predicted segments with the actual objects.

4.3. Results Overview

Table 1 shows GroupViT's performance on PASCAL VOC 2012, comparing it to traditional supervised models. Despite using only text supervision, GroupViT achieved competitive results.

Impact of Group Tokens: The number of group tokens used has a direct impact on the segmentation quality, as shown in Table 2.

Table 2. Impact of Group Tokens

Group Tokens	mIoU (%)
16 Tokens	28.6
64 Tokens	39.3

Comparison with Fully-Supervised Models: Table 3 compares GroupViT's zero-shot performance with fully-supervised ViT models.

Table 3. Comparison with Fully-Supervised Models

Model	Method	mIoU (%)
ViT (Supervised)	Pixel-Level Labels	53.0
GroupViT (Ours)	Zero-Shot	52.3

Ablation Study Results: The impact of assignment strategies is highlighted in Table 4.

Table 4. Ablation Study Results

Method	mIoU (%)
Soft Assignment	12.0
Hard Assignment	36.7

Comparison with Zero-Shot Baselines: Table 5 compares GroupViT with other zero-shot segmentation baselines.

Table 5. Comparison with Zero-Shot Baselines

Model	Approach	mIoU (%)
ViT	Pixel-Wise Classification	20.1
ViT	K-means Clustering	25.0
GroupViT (Ours)	Text Supervision	52.3

5. Discussion

The experimental results underscore GroupViT's potential as a robust segmentation model that does not require pixel-level labels. Key takeaways include:

5.1. Strengths

Generalization: GroupViT's zero-shot capability allows it to segment unseen object categories accurately, relying solely on text descriptions. This adaptability is crucial for dynamic environments and evolving datasets.

Flexibility: The hierarchical grouping mechanism enables GroupViT to segment objects of various shapes and sizes, outperforming traditional grid-based approaches in complex scenes.

5.2. Limitations

Background Complexity: The model struggles with highly complex backgrounds, where text descriptions may be insufficient to differentiate subtle visual elements.

Misclassification Risks: Ambiguous or vague text descriptions can lead to misclassifications, highlighting the need for more precise and detailed training data.

5.3. Future Directions

Enhanced Grouping Techniques: Integrating advanced techniques like pyramid pooling or dilated convolutions could improve the model's ability to handle detailed and complex scenes.

Expanding Training Datasets: Utilizing larger, more diverse image-text datasets could enhance GroupViT's robustness, particularly for handling zero-shot segmentation in unfamiliar contexts.

References

[1] Dosovitskiy, A., et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." In *International Conference on Learning Representations (ICLR)*, 2021.

[2] Radford, A., et al. "Learning Transferable Visual Models from Natural Language Supervision." In *International Conference on Machine Learning (ICML)*, 2021.

[3] Everingham, M., et al. "The PASCAL Visual Object Classes (VOC) Challenge." *International Journal of Computer Vision (IJCV)*, 2010.

[4] Zhang, H., et al. "GroupViT: Semantic Segmentation Emerges from Text Supervision." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[5] Chen, L.-C., et al. "Rethinking Atrous Convolution for Semantic Image Segmentation." In *arXiv preprint arXiv:1706.05587*, 2017.

[6] Carion, N., et al. "End-to-End Object Detection with Transformers." In *European Conference on Computer Vision (ECCV)*, 2020.

[7] He, K., et al. "Mask R-CNN." In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[8] Lin, T.-Y., et al. "Microsoft COCO: Common Objects in Context." In *European Conference on Computer Vision (ECCV)*, 2014.

[9] Xie, E., et al. "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers." In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[10] Liu, Z., et al. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows." In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[11] Li, X., et al. "Expectation-Maximization Attention Networks for Semantic Segmentation." In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[12] Zheng, S., 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 (CVPR)*, 2021.

[13] Kirillov, A., et al. "Panoptic Segmentation." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[14] Long, J., et al. "Fully Convolutional Networks for Semantic Segmentation." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[15] Zhou, B., et al. "Semantic Understanding of Scenes through ADE20K Dataset." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[16] Chen, L.-C., et al. "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation." In *European Conference on Computer Vision (ECCV)*, 2018.

[17] Cordts, M., et al. "The Cityscapes Dataset for Semantic Urban Scene Understanding." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[18] Wang, H., et al. "Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation." In *European Conference on Computer Vision (ECCV)*, 2020.

[19] Huang, Z., et al. "CCNet: Criss-Cross Attention for Semantic Segmentation." In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.