Zero-shot semantic segmentation

Tarun Kumar, Diganta Diasi, Major Vaibhav Mishra, Anjali Jaiswal and Aman Kumar Prasad Gond

***Abstract*—This project aims to develop a novel model for zero-shot semantic segmentation (ZS3), a challenging computer vision task. ZS3 involves segmenting objects in an image into different categories without prior training on those specific categories. Our goal is to create a model that leverages both visual and auxiliary information to achieve accurate and fine-grained segmentation, addressing the limitations of existing approaches. This work has applications in various fields and is valuable when obtaining labeled data for all possible classes is impractical or costly. By bridging the gap between seen and unseen classes using semantic information, we aim to contribute to the advancement of computer vision technology.**

I. Introduction

A. Background

Zero-Shot Learning: Zero-shot learning is an approach where a model is trained to classify data into classes it hasn't seen during training, using auxiliary information like attributes or semantic embeddings to bridge the gap between seen and unseen classes.

Semantic Segmentation: Semantic segmentation, aims to assign a corresponding and unique class label to each pixel in an image, indicating what is being represented by that pixel. This task is also known as dense prediction, since we are predicting for each pixel in the image.

In semantic segmentation, an image is divided into meaningful and semantically coherent regions, where each pixel is assigned a class label that corresponds to the object or category it belongs to. Unlike instance segmentation, which distinguishes between individual instances of the same class, semantic segmentation focuses on classifying each pixel into a broad category without distinguishing between different instances of that category.

Semantic segmentation can be a useful alternative to object detection because it allows the object of interest to span multiple areas in the image at the pixel level. This technique cleanly detects objects that are irregularly shaped, in contrast to object detection, where objects must fit within a bounding box

Because semantic segmentation labels pixels in an image, it is more precise than other forms of object detection. This makes semantic segmentation useful for applications in a variety of industries that require precise image maps.

B. Challenges and Motivation

Zero-Shot Semantic Segmentation: The goal is to segment objects or regions in an image into different classes or categories without any prior training on those specific categories.
This means that the model has to recognize and segment objects it has never seen during training.

1) To achieve zero-shot semantic segmentation, techniques often leverage external information or embeddings that describe object categories.
   For example, you can use word embeddings like Word2Vec or pre-trained models like BERT to associate textual descriptions with object categories. Then, this information can be combined with the image features to perform segmentation.
2) It requires handling challenges such as handling unseen categories and dealing with the semantic gap between textual descriptions and visual features.

It has applications in many fields like autonomous driving, augmented Reality, robotic vision, remote sensing, medical imaging. Zero-shot learning is valuable in situations where it's impractical or costly to obtain labeled data for all possible classes, such as in fine-grained object recognition, natural language processing, and various other domains. It leverages semantic information to bridge the gap between seen and unseen classes.

II. Related Works

A. *Delving into Shape-aware Zero-shot Segmentation*

Shape-Aware Zero-Shot Semantic Segmentation (SAZS) is a novel framework for zero-shot semantic segmentation. It enforces vision-language alignment during training and aligns predicted semantic region boundaries with ground truth regions. SAZS leverages Laplacian matrix eigenvectors to decompose inputs into eigensegments, which are fused with learning-based predictions.

The paper addresses drawbacks with earlier models include challenges in open-world applications, underexploration of dense prediction tasks with large-scale pretrained models, and issues with fine shape delineation. The paper presents experimental results demonstrating SAZS's effectiveness on benchmark datasets, surpassing previous state-of-the-art methods.

However, SAZS has limitations, including its reliance on specific training data for vision-language and boundary alignment, potential inaccuracies in segmenting complex shapes, and variations in performance across different datasets and scenarios.

B. *Context-aware Feature Generation for Zero-shot Semantic Segmentation*

The paper introduces CaGNet, a novel technique for zero-shot semantic segmentation, addressing the challenge of segmenting objects without dense pixel-level annotations. CaGNet leverages semantic word embeddings to transfer knowledge between object categories.

CaGNet incorporates a Contextual Module (CM) into the segmentation network to capture pixel-wise contextual information, which extends beyond spatial arrangement to include factors like object placement and background context. The CM takes semantic word embeddings and pixel-wise contextual latent codes, providing real features and contextual codes for all pixels. It dynamically weighs different contextual information scales for pixels, resolving feature generation ambiguity.

The combination of pixel-wise contextual latent coding and semantic word embeddings improves feature generation. The paper claims that CaGNet achieves state-of-the-art results on three benchmark datasets for zero-shot segmentation, demonstrating its effectiveness in tackling this challenging task.

C. *Decoupling Zero-Shot Semantic Segmentation*

The ZS3 approach faces two primary challenges. Firstly, transferring knowledge from known to unknown classes, often relying on language models pre-trained on text data, limiting their performance in visual tasks. Second, establishing correlations between pixel-level visual features and semantic features, which is unnatural since humans use words or texts to describe objects.

To address these challenges, the paper proposes a decoupling approach, dividing the procedure into class-agnostic grouping and segment-level zero-shot classification and introduces the ZegFormer model, a zero-shot semantic segmentation model. ZegFormer utilizes a transformer decoder for generating segment-level embeddings and incorporates class-agnostic grouping (CAG) and segment-level zero-shot classification (s-ZSC). It employs mask projection for mapping embeddings to binary mask predictions and semantic projection for relating segment-level embeddings to semantic features from a pre-trained text encoder.

Extensive experiments confirm the effectiveness of the ZegFormer model on various zero-shot segmentation benchmarks, surpassing existing methods by a significant margin in terms of mean Intersection over Union (mIoU) for unseen classes.

However, the ZegFormer model is not without drawbacks. Decoupling its components may lead to compatibility issues between different parts of the system, demanding complex integration efforts. Moreover, the process can increase the model's complexity and training pipeline, potentially requiring additional modules or steps to manage unseen object classes, making the system more intricate to handle and maintain.

D. *ZS3*

The paper introduces ZS3Net, a novel architecture for zero-shot semantic segmentation, combining a deep visual segmentation model with semantic word embeddings to understand and segment object classes based on textual descriptions. It also proposes zero-shot benchmarks on standard datasets like Pascal-VOC and Pascal-Context and emphasizes addressing pixel classification for object categories without training examples. However, the paper has potential limitations, such as data dependency on the availability and quality of textual descriptions and word embeddings, uncertainty about how well the model generalizes to real-world scenarios, and limited focus on adapting to entirely novel objects beyond the initial training set.

E. *ZegClip*

In this work, the authors aim to simplify the process of zero-shot semantic segmentation by extending CLIP's zero-shot capability from image-level to pixel-level, eliminating the need for a two-stage approach that involves proposal generation and image-level classification. The paper presents an efficient one-stage solution called ZegCLIP.

The primary idea is to use a lightweight decoder to match text prompts with local embeddings from CLIP. However, this initial approach tends to overfit to seen classes and struggles with unseen classes. To address this, the authors propose three key modifications:

1) Deep Prompt Tuning (DPT): Instead of fine-tuning or fixing the CLIP image encoder, they use DPT. This helps avoid overfitting to seen classes and retains CLIP's zero-shot capacity.
2) Non-mutually Exclusive Loss (NEL): The authors apply an NEL function for pixel-level classification, generating posterior probabilities independent of other class logits.
3) Relationship Descriptor (RD): The major innovation is the introduction of an RD to incorporate image-level priors into text embeddings. This helps prevent overfitting to seen classes and enhances generalization.

By incorporating these modifications, the authors create ZegCLIP, a straightforward yet effective zero-shot semantic segmentation model. ZegCLIP demonstrates superior performance on public benchmarks, outperforming state-of-the-art methods by a significant margin in both "inductive" and "transductive" zero-shot settings. Furthermore, compared to the two-stage method, ZegCLIP achieves a remarkable five-fold speedup during inference.

Each of these papers or approaches addresses zero-shot semantic segmentation using different techniques

and models, and they have their respective strengths and potential limitations.

III. Proposed Work

We propose a method for zero-shot segmentation which represents a complex yet intriguing approach to image understanding. It brings together a variety of techniques to discern and segment objects without the need for specific class training. The process involves concurrently feeding the input image into a visual encoder and a shape delineator. This dual-path approach is then followed by spectral decomposition, utilizing singular value decomposition (SVD), which allows us to extract essential spectral information.

A. Block Diagram

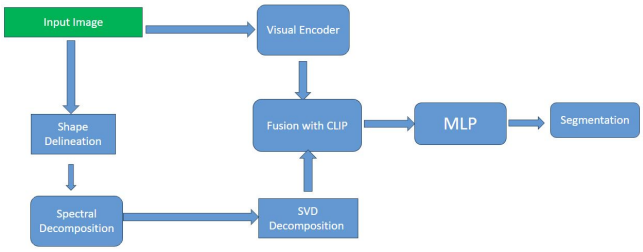

Fig. 1. The outputs from both the visual encoder and SVD are seamlessly integrated via CLIP (Contrastive Language-Image Pretraining), creating a unified space where visual and spectral features harmoniously coexist. This shared representation is then directed into a Multi-Layer Perceptron (MLP) to carry out zero-shot segmentation, effectively classifying objects even when they haven't been encountered during the training phase.

The strength of our methodology lies in its comprehensive fusion of visual, shape, and spectral data. This fusion empowers our approach to potentially segment previously unseen classes, demonstrating a sophisticated yet adaptable algorithmic pipeline for image analysis and classification in a wide range of real-world scenarios.

However, we faced significant challenges in implementing Zegclip and we were not able to implement our proposes idea in limited time frame.

IV. Experimental Details

A. Datasets

Experimental evaluation is done on the two datasets: Pascal-VOC 2012 and Pascal-Context.

Pascal-VOC contains 1,464 training images with segmentation annotations of 20 object classes. We have adopted additional supervision from semantic boundary annotations during training.

Pascal-Context provides dense semantic segmentation annotations for Pascal-VOC 2010, which comprises 4,998 training and 5,105 validation images of 59 object/stuff classes.

B. Training Details

Currently, we are in the preliminary phase of translating our proposed solution into functional code. As we delve into this process, we are continuously exploring various strategies to effectively advance our solution. In our endeavor to gain comprehensive insights into past models and their real-world applications, we undertook the implementation of two specific models, ZS3net and ZegCLIP, on the Pascal-VOC dataset. This initiative was aimed at fostering a deeper understanding of the models' practical implications within the context of our project.

Our initial implementation efforts were focused on the ZS3Net model, and the following table presents the results of this endeavor. Notably, for the baseline model, we referenced the results from the ZS3 paper. However, our findings diverged from those reported in the research paper. The discrepancy arose due to the unavailability of the SBD dataset, necessitating code modifications to enable exclusive compatibility with the Pascal-VOC dataset. It's important to note that the research paper also incorporated the COCO dataset, which, unfortunately, couldn't be accommodated by our CPU desktop during the training phase.

Nevertheless, we present the results obtained from the Pascal-VOC dataset for 2 and 4 classes in the table below.

TABLE I
Performance metrics for K model in Seen and Unseen categories (ZS3Net).

K	Models	Seen			Unseen		
		PA	MA	mIoU	PA	MA	mIoU
2	Baseline	92.1	79.8	68.1	11.3	10.5	3.2
	ZS3Net	90.6	79.9	64.0	52.8	53.7	35.4
4	Baseline	89.9	72.6	64.3	10.3	10.1	2.9
	ZS3Net	92.0	78.3	66.4	43.1	45.7	23.2

Throughout our project, we dedicated efforts to incorporate the ZegClip model, yet encountered several obstacles along the way. One of the major hurdles was the reliance on an outdated version of the mmcv library, rendering it incompatible with the present system requirements. Consequently, attempts to build mmcv wheels using the GPU were unsuccessful, compelling us to resort to CPU usage, which, unfortunately, did not resolve our persistent issues with the mmcv library. In particular, version 1.4.4 of mmcv posed significant challenges, with the program repeatedly throwing an error indicating the absence of the "mmcv._ext" module.

Confronted with this setback, we embarked on a comprehensive examination of the script files and the 'train.py' code, determined to implement the necessary updates to rectify this predicament and ensure seamless compatibility with the latest versions of the mmcv library. Despite making notable headway and achieving successful execution

during the training phase, we are still grappling with complications during the testing phase of the code.

C. Baseline Methods

Our methodology is based on a basic model that is quite similar to the two-stage OpenAI Clip model implementation. Still, our model brings about some changes, mainly by using the idea of Singular Vector Decomposition. We demonstrate our commitment to pushing the limits of semantic segmentation research by applying this novel approach in the context of Zero-Shot Semantic Segmentation using patches. In order to comprehend the workings of the CLIP model within the framework of the Zero-Shot Semantic Segmentation (ZS3) task, it is imperative to take into account the underlying mechanism. Recent research efforts have addressed the application of CLIP in zero-shot segmentation problems by proposing a two-stage paradigm, given its remarkable zero-shot classification capabilities as shown in large-scale pre-training. In this paradigm, a class-agnostic generator is trained in the first step, and in the second stage, CLIP is used as a zero-shot image-level classifier. The [cls] token of each proposal and text embeddings are compared for similarity in order to classify the proposals. This method has shown to be successful, but because of the two separate image encoding procedures required by its design, it has a large computing overhead and presents a substantial challenge in terms of computational efficiency and resource allocation.

D. Evaluation Metrics

In the context of our experiments, we have meticulously adopted a set of standard semantic segmentation metrics, namely pixel accuracy (PA), mean accuracy (MA), and mean intersection-over-union (mIoU). Our approach aligns closely with the methodology outlined in the Zero-Shot Semantic Segmentation paper, wherein we also place significant emphasis on the harmonic mean (hIoU) of both seen and unseen mIoUs during our analysis. The deliberate choice of using the harmonic mean over the traditional arithmetic mean stems from our recognition of the potential bias arising from seen classes that often exhibit significantly higher mIoU values, potentially overshadowing the overall evaluation.

By incorporating the harmonic mean, we strive to provide a more comprehensive and nuanced assessment, allowing us to effectively gauge the performance of the ZS3 models across different classes. This nuanced metric enables us to capture the interplay between seen and unseen classes, offering deeper insights into the model's capacity to generalize and perform well across various semantic segmentation tasks.

V. Results

A. Comparison with State-of-the-art Methods

1) Quantitative Results: In our current developmental phase, it's important to clarify that the actual implementation of our proposed methodology on a real-world dataset is still underway, as we are primarily focused on laying the foundational groundwork for the model's code. However, we have taken strides in the practical domain by endeavoring to apply well-established models, including ZS3 and ZegClip, to the intricate Pascal VOC dataset. Notably, our experiments encompassed the execution of these models across both two and four classes, requiring us to make crucial refinements to the original ZS3 code provided by its creators. This undertaking was prompted by the unavailability of the SBD dataset, necessitating adaptations to ensure the smooth operation of the models in our specific context.

Despite grappling with certain setbacks arising from compatibility issues, particularly pertaining to the outdated mmcv module in the case of the ZegClip model, our persistence yielded encouraging and promising results, especially evident in the performance outlined in Table 1 for the ZS3 model. These initial findings underscore the potential efficacy of our proposed approach within the complex realm of semantic segmentation, setting a strong foundation for further advancements in our ongoing research endeavors.

VI. Conclusion

In conclusion, our approach to zero-shot segmentation is complex and promising. It integrates various techniques to segment objects without specific class training. This involves using a visual encoder and a shape delineator in tandem, followed by spectral decomposition through SVD for essential spectral information. The model is comprehensive fusion of visual, shape, and spectral data, enabling potential segmentation of unseen classes. It offers a sophisticated and adaptable pipeline for image analysis and classification in real-world scenarios.

However, implementing ZegClip posed significant challenges, preventing us from fully realizing our proposed idea within the project's time constraints.

In our experiments, we focused on datasets like Pascal-VOC 2012 and Pascal-Context, with added supervision from semantic boundary annotations during training. Our preliminary work in translating the proposed solution into functional code provided insights into models like ZS3net and ZegCLIP, enhancing our understanding of their practical implications in our project context.

Despite initial implementation challenges and dataset discrepancies, we are actively addressing issues, and our project remains ongoing.