CPSeg: Finer-grained Image Semantic Segmentation via Chain-of-Thought Language Prompting

Lei Li
University of Copenhagen

lilei@di.ku.dk

Abstract

Natural scene analysis and remote sensing imagery offer immense potential for advancements in large-scale language-guided context-aware data utilization. This potential is particularly significant for enhancing performance in downstream tasks such as object detection and segmentation with designed language prompting. In light of this, we introduce the ***CPSeg*** *(****C****hain-of-Thought Language* ***P****rompting for Finer-grained Semantic* ***Seg****mentation), an innovative framework designed to augment image segmentation performance by integrating a novel "Chain-of-Thought" process that harnesses textual information associated with images. This groundbreaking approach has been applied to a flood disaster scenario.* ***CPSeg*** *encodes prompt texts derived from various sentences to formulate a coherent chain-of-thought. We use a new vision-language dataset, FloodPrompt, which includes images, semantic masks, and corresponding text information. This not only strengthens the semantic understanding of the scenario but also aids in the key task of semantic segmentation through an interplay of pixel and text matching maps. Our qualitative and quantitative analyses validate the effectiveness of* ***CPSeg****.*

Image segmentation has emerged as a critical component in the analysis of remote sensing imagery, aiming to partition an image into multiple segments, or sets of pixels, that correspond to distinct objects or object components [3, 13, 16, 30, 34, 36]. Its significance is further magnified in the context of remote sensing as a global observation system, encompassing various applications such as urban planning, resource management [20, 25, 35], environmental monitoring [14, 15, 40], and particularly, disaster response [23]. However, accurately segmenting remote sensing images using language modeling is confronted with complexities arising from diverse textures, irregular shapes, corresponding alignment, and varying scales present in these images. Consequently, the development of effective segmentation with contrastive VL (Vision-Language) pre-training remains an ongoing challenge and a research priority.

Figure 1. In the visualization of fine-grained image semantic segmentation, the incorporation of chain-of-thought language prompting proves instrumental in attaining meticulous and accurate outcomes. The left panel of the illustration showcases a bird's-eye view, juxtaposed with the output mask derived from the application of CPSeg to the original image. This arrangement offers a comprehensive outlook on the original image and its corresponding segmentation. Conversely, the right panel emphasizes the chain-of-thought prompting process, employing diverse question types. It showcases color-coded masks that represent each thought-provoking question in the chain, accompanied by their finer-grained segmentation outputs. This dual-paneled visualization approach effectively portrays the intricate procedure and intricate outcomes of image semantic segmentation through the utilization of chain-of-thought language prompts.

Drawing inspiration from the success of contrastive VL pre-training, specifically CLIP [22], several recent works [12, 24, 32, 33] have explored CLIP-based segmentation approaches to improve the transfer of language features and enhance segmentation performance. VL segmentation

methods [11,12,32] also face challenges in terms of training with additional image data, acquiring segmentation annotations, or obtaining natural language supervision. These challenges are crucial when adapting pre-trained vision-language models to downstream segmentation tasks, and they significantly impact the optimization and scalability of these models.

Existing methodologies for VL based image semantic segmentation often overlook the incorporation of human cognitive processes and sequential thought patterns [29], particularly within the field of remote sensing where such approaches are scarce. The work of Wei et al. [29] introduced a chain-of-thought network that demonstrated the effectiveness of this approach in natural language processing. However, there has been limited exploration of this methodology in the context of image segmentation. This gap in research becomes evident when analyzing complex images, such as flood scenes [23], as depicted in Figure 1. A human observer naturally engages in a tiered process of analysis, initially identifying distinct classes within the image (e.g., buildings or roads), followed by evaluating the quantity and extent of their impact due to flooding (e.g., number of submerged buildings or impassable roads). Unfortunately, this sequential and logical cognitive process remains largely unexplored in existing image semantic segmentation frameworks. This presents an opportunity to enhance these models by integrating insights from human cognitive processing.

Meanwhile, the progress in remote sensing technologies has opened up intriguing possibilities for enhancing image segmentation methodologies. In this study, we aim to investigate this potential by introducing a novel framework for image segmentation that harnesses language-guided context-aware data, an approach that has been widely utilized in the analysis of natural scenes but remains relatively unexplored in the domain of remote sensing imagery. Our experimental results have substantiated the effectiveness of the proposed framework, specifically in its utilization of a chain-of-thought process to iteratively incorporate textual information into the image segmentation pipeline.

Our work explore to address this gap by introducing a novel framework that utilizes a chain-of-thought continual-vision strategy to enhance image segmentation in remote sensing. While several studies have explored different strategies to improve image segmentation accuracy, our approach focuses on leveraging the chain-of-thought process, which involves sequential reasoning to analyze complex images. By mimicking human cognition, this approach enables the sequential thought process that humans employ to identify, relate, and understand various elements within an image. Our framework integrates textual information derived from images in a continuous manner, effectively leveraging language data to enhance segmentation performance. This approach is particularly beneficial in time-critical scenarios such as disaster response, where the ability to not only identify but also understand the spatial relationships and context of objects through chain-of-thought processing in an image can be crucial.

Our findings demonstrate that our proposed method significantly improves segmentation outcomes in a flood disaster scenario when using the FloodPrompt for empirical validation. The utilization of a text encoder to process prompt texts from different sentences, and the incorporation of the encoded information in the semantic segmentation task, have proven to be particularly advantageous. Our contributions can be summarized as follows:

- We introduce a novel methodology for finer-grained image semantic segmentation specifically designed for remote sensing imagery, harnessing the power of language prompting.
- We propose a novel task that incorporates the concept of chain-of-thought prompting into the domain of image semantic segmentation, paving the way for more advanced segmentation algorithms.
- Through an extensive validation process, we demonstrate that our proposed FloodPrompt dataset outperforms conventional methods in terms of label semantic segmentation and language-guided approaches, highlighting its superior efficacy.

We begin this paper with a comprehensive introduction, which includes a detailed overview of our novel framework and a review of related work in the field. We then proceed to present a thorough explanation of our methodology, outlining the various components and their roles. Following this, we provide an extensive analysis of our experimental results, encompassing both quantitative and qualitative assessments. To further elucidate the contributions of different components, we also include an ablation study. Finally, we conclude the paper with a discussion on the implications of our findings and outline potential avenues for future research.

1. Related Work

Vision-language pre-training. Recent advancements in vision-language models, which are pre-trained on large-scale image-text datasets, have demonstrated remarkable efficacy in adapting to novel tasks within the context of zero-shot and few-shot learning, spanning diverse domains [1, 2, 31]. These developments highlight the significant potential for applying such models in complex computational domains. Notably, models incorporating dual encoders, multi-modal encoders, and encoder-decoders, such as CLIP, have further improved cross-modal representation

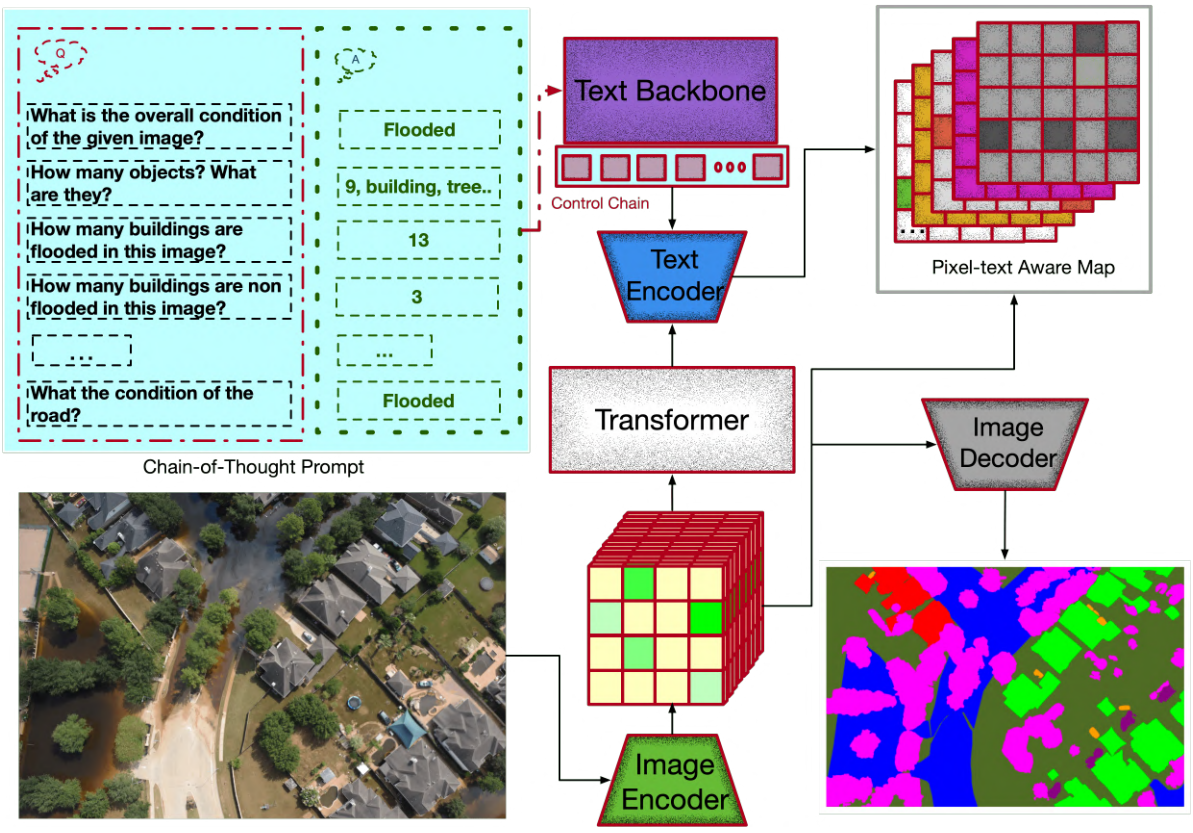

Figure 2. The CPSeg framework is designed for image semantic segmentation and involves several key steps. Firstly, it extracts embeddings from both images and chain-of-thought text prompts. These embeddings are then used to compute pixel-text aware maps, which represent a novel adaptation of CLIP's image-text matching problem. CPSeg introduces a context-aware approach that enables dense prediction in segmentation tasks. Additionally, it incorporates a transformer module [27] for both the text and vision backbone, leveraging pre-trained knowledge to improve the accuracy of segmentation results. This innovative methodology expands the possibilities for achieving more precise and nuanced image semantic segmentation.

quality through contrastive pre-training. Concurrently, the "pre-training + fine-tuning" paradigm has revolutionized computer vision and natural language processing by initially pre-training models on extensive datasets like ImageNet [10], JFT [26], and Kinetics [6], followed by fine-tuning for various downstream tasks. This framework often evolves into a prompt-based paradigm, wherein downstream tasks are reformulated to align with those addressed during the pre-training process. Collectively, these strategies reflect the evolving landscape of vision-language pre-training, holding promise for advancing performance in complex tasks.

Image segmentation with Vision-language. Image segmentation [7, 17, 19, 37] remains a central yet challenging task in computer vision, particularly when segmenting novel visual categories. A variety of approaches have been explored, including unsupervised, zero-shot segmentation, and methods leveraging vision-language models. Unsupervised segmentation approaches often focus on clustering dense image representations and matching these to corresponding segmentation categories, while vision-language (VL) [18, 39] driven strategies aim to replace the matching process with text encoders for enhanced efficiency and transferability. Meanwhile, transferring methods [4] often necessitate class-agnostic or class-specific segmentation annotations, despite recent innovations using VL models. In the context of these developments, this paper explores image-free semantic segmentation, aiming for practical applicability in scenarios where only segmentation vocabulary is given, providing a simpler alternative to collecting images or other annotations. To address the significant annotation burden associated with previous supervised pre-training settings, several self-supervised pre-training approaches have been introduced in the field of dense prediction [5, 24, 28] with fine-tuning strategy that harnesses the knowledge embedded within large-scale vision-language pre-trained models. Importantly, this strategy incorporates language infor-

mation as a guiding component within the learning process, marking a distinct departure from traditional methodologies. The evolving intersection of computer vision and natural language processing fields, especially with vision-language pre-training, offers new perspectives for these challenges, with models like CLIP demonstrating impressive transferability over diverse classification datasets. Yet, very few attempts have been made to apply such models to image segmentation prediction tasks, making it a compelling area for future exploration.

2. Methodology

The CPSeg begins with the extraction of embeddings from both the input image and the chain-of-thought text prompts. These embeddings are then utilized to generate pixel-text aware maps. we adapt the chain-of-thought process to suit the specific requirements of our task, namely, improving the interpretative capability of our model with the help of corresponding textual information.

2.1. Overview

The CPSeg framework incorporates a dynamic calculation of the pixel-text match loss within the pixel-text aware map, which is updated by the transformer's parameters as more prompts are provided. The resulting score maps are then fed into a decoder that utilizes ground-truth labels for supervision. Moreover, CPSeg capitalizes on the wealth of pre-trained knowledge(CLIP) by leveraging contextual information present in images to guide the language model prompts. This is achieved through the integration of a transformer module, enabling the framework to optimize its understanding of the image's context and thereby improving the quality of the segmentation results. The overall architecture of CPSeg establishes it as a robust and efficient tool for image semantic segmentation.

The chain-of-thought process as show in Figure 3, a crucial element of this framework, is rooted in the theoretical understanding of human cognition. The methodology of CPSeg employs a pre-trained Vision Transformer, denoted as V. It mandates various prerequisites including the total number of tasks, T; a comprehensive training set, denoted as $\{(I_i^t, L_i^t)\}_{i=1,t=1}^{N_t,T}$; a collection of prompts, Q, and their respective keys, K. Subsequent steps involve the determination of specifics like the number of training epochs for the t-th task, E_t, the learning rate η, and a balancing parameter λ. The objective is to update and optimize the parameters of V, Q, and K, using a hierarchical prompting mechanism. For more details about prompting mechanism, refer to supplement.

It breaks down complex tasks into a sequence of smaller, more manageable decisions. This offers a practical approach for handling intricate data analysis tasks such as image segmentation. Formally, given an image I with pixel data $P = p_1, p_2, \ldots, p_n$, the chain-of-thought process handles the segmentation task as a sequence of decisions concerning each pixel p_i. CPSeg enhances this process by integrating context-aware data guided by language. In particular, we construct a chain of thoughts $C = c_1, c_2, \ldots, c_m$, where each thought c_i corresponds to a sentence s_i in the text accompanying the image. Each thought c_i consists of a text encoder $T(c_i)$ and a pixel-level segmentation function $f_{c_i}(p_i)$, generating a sequence of segmentation decisions $D = d_1, d_2, \ldots, d_m$, with each decision d_i corresponding to a thought c_i.

To support this process, we employ a text encoder E that generates encoded representations from diverse sentence prompts. Formally, for a given sentence s_i, we derive its encoded representation $e_i = E(s_i)$, where e_i captures the semantic details in s_i. This encoding procedure enables our framework to leverage the semantic context provided by the language data, enhancing the capability of the chain-of-thought process. The encoded data subsequently facilitates the downstream task of semantic segmentation. Each segmentation function $f_{c_i}(p_i)$ in the thought c_i employs the corresponding encoded representation e_i to make better-informed decisions about pixel classification. Specifically, for a given pixel p_i, the segmentation decision d_i is computed as $d_i = f_{c_i}(p_i, e_i)$. This allows our framework to make use of both the spatial information from the pixels and the semantic information from the text, achieving more accurate and context-aware segmentation outcomes through chain-of-thought prompting. The Figure 3 illustrates the flow of the chain-of-thought prompting process in the CPSeg framework.

2.2. Language Prompting.

The proposed framework begins with an initial pre-trained segmentation network S. The system is designed to handle T tasks, each consisting of its own set of epochs denoted as E_t. At each epoch, a mini-batch is randomly sampled from the dataset, and suitable prompts and keys are generated for each image within the mini-batch. The algorithm incorporates a control mechanism that verifies the relevance of the query questions and corresponding answers. For instance, if the question relates to the number of flooded buildings, the controller specifically seeks out this information and retrieves the appropriate answer.

The framework supports various types of prompts, including simple counting and condition recognition, as depicted Figure 3. These prompts and keys are applied to the segmentation network, and the loss is calculated. Subsequently, the segmentation network is refined using gradient descent. This process is repeated for all images, mini-batches, and tasks, resulting in an updated segmentation network.

The construction of the entire chain-of-thought is hier-

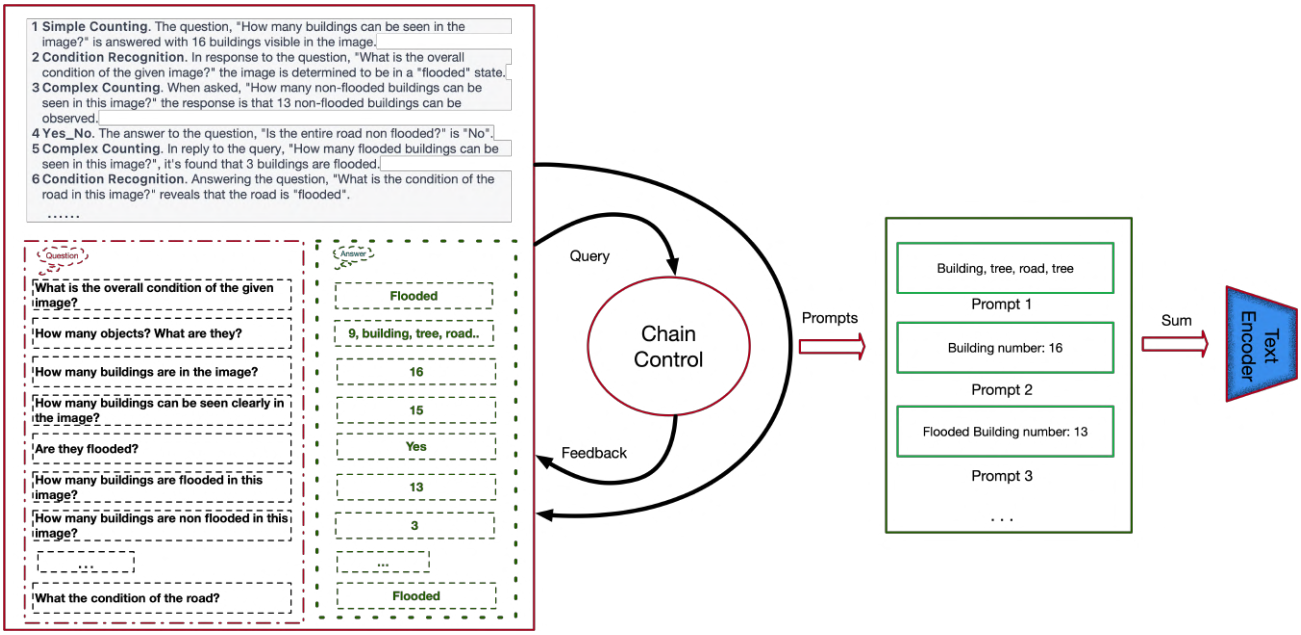

Figure 3. Our chain-of-thought prompting pipeline is intricately designed to elicit comprehensive responses. When relevant classes are identified, the pipeline proceeds to inquire about the precise numbers involved. Furthermore, it thoroughly examines all the potential questions within the context, ensuring a robust and detailed understanding.

archical in nature, starting with macro-level considerations before delving into the specifics. In curating the prompts, we've orchestrated the keys in a hierarchical manner, commencing from overarching descriptors of the entire image and subsequently delving into more granular details, such as the presence of infrastructure elements like buildings and roads. This structured approach ensures a systematic progression from a macroscopic view to refined prompting. In the event of flood-related incidents, subsequent prompts elucidate on the inundation of infrastructure elements, quantifying inundated buildings, and further delineating the intricacies of the flooding scenario and its concomitant implications.

For instance, the initial query might be regarding the presence of flooding, which then transitions into contemplation about the presence of buildings, and subsequently, the number of buildings inundated.

2.3. Vision Backbone.

The CPSeg framework tackles the challenge of semantic segmentation by adopting a Vision-and-Language (VL) encoder-decoder model. The objective is to decode a category word for each densely populated image region, considering M semantic categories of interest. However, a key challenge arises when specific semantic category words are tokenized into multiple subwords within the dictionary, introducing complexity to the task.

To overcome this challenge, we employ a Vision Transformer (ViT) as the encoder, enabling the extraction of highly detailed and efficient visual content representations. During the inference process, the encoder and decoder interact synergistically to generate the semantic segmentation mask. The decoder plays a vital role in converting the dense feature representation obtained from the encoder into category predictions for each image region. It handles the complexities arising from semantic relationships and potential subword tokenization. This approach strikes a balance between theoretical complexity and practical efficiency, providing an effective methodology for semantic segmentation tasks with reduced complications.

2.4. Loss fuction.

Pixel-Text Matching Loss. The alignment between the image pixels and textual prompts can be quantified using the Pixel-Text Matching Loss function, mathematically denoted as L_{PTM}. Let us denote the set of image pixels as $P = p_i i = 1^N$ and the set of textual prompts as $T = t_j j = 1^M$. A similarity score $s(p_i, t_j)$, computed using an appropriate metric such as the cosine similarity or dot product, is assigned to each pair of pixel p_i and text prompt t_j. The Pixel-Text Matching Loss is then computed as the negation of the accumulated similarity scores:

$$L_{\text{PTM}} = -\sum_{i=1}^{N}\sum_{j=1}^{M} s(p_i, t_j). \quad (1)$$

This loss function is minimized during the model train-

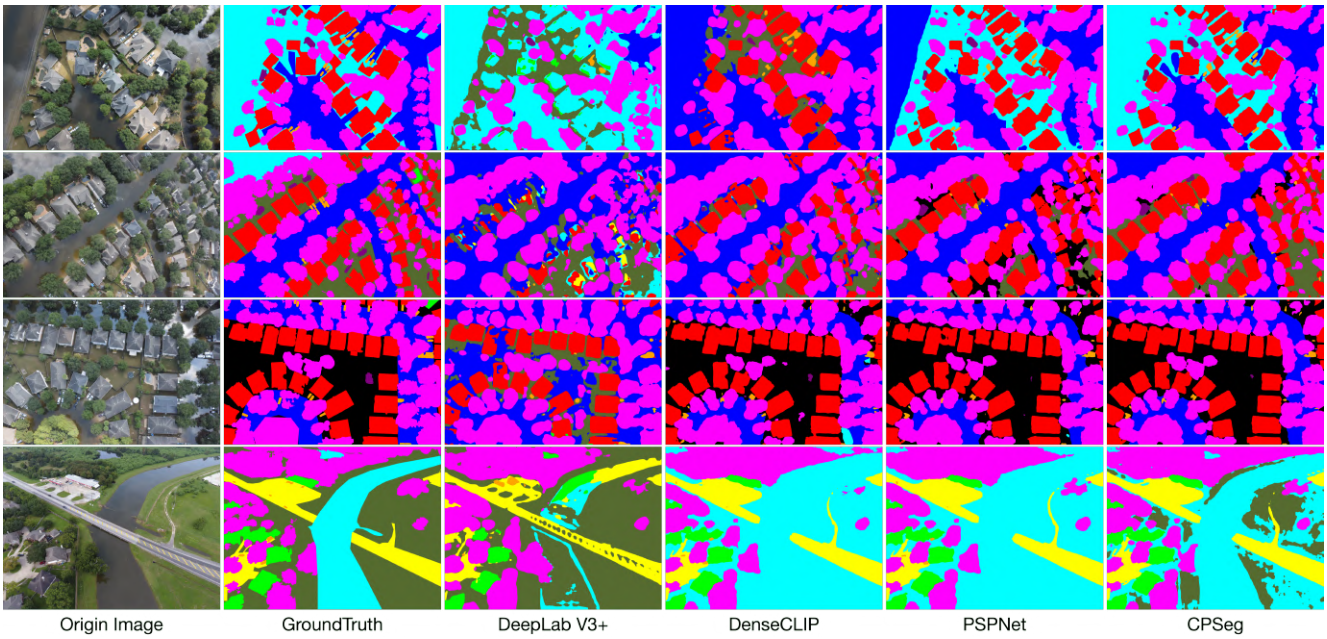

Figure 4. Visual Comparison of Segmentation Results: A comparative study of segmentation performance across DeepLab V3+, DenseCLIP, PSPNet, and our proposed CPSeg.

ing phase, with the aim of maximizing the overall similarity between the pixels and prompts. By doing so, the model is incentivized to learn representations that improve the correspondence between similar pixels and prompts.

Semantic Segmentation Loss. Our method compute score maps in segmentation. These score maps $\mathbf{s} \in \mathbb{R}^{H_4 W_4 \times K}$ can be treated as smaller-scale segmentation results. We compute a segmentation loss on them:

$$\mathcal{L}_{\text{seg}} = \text{CrossEntropy}(\text{Softmax}(\mathbf{s}/\tau), \mathbf{y}), \qquad (2)$$

where $\tau = 0.07$ is a temperature coefficient following prior work [39], and $\mathbf{y} \in 1, \ldots, K^{H_4 W_4}$ denotes the ground truth labels.

3. Experiments

3.1. Dataset

To validate the efficacy of our proposed methodology, we adapted FloodNet dataset [23] for FloodPrompt and utilized FloodPrompt for our experiments. FloodPrompt is a diverse and challenging dataset, containing a variety of remote sensing images pertinent to flood scenarios. Given the paucity of studies addressing image segmentation in such scenarios, we propose FloodPrompt provides a relevant and complex testing ground for our framework. The dataset encompasses numerous instances of flooding events captured through remote sensing technology, all annotated with detailed text descriptions, making it a suitable candidate for evaluating our language-guided segmentation approach.

The textual descriptions associated with each image were preprocessed, tokenized, and encoded using the text encoder component of our framework. To ensure a fair comparison, the proposed method was compared with state-of-the-art segmentation methods, under identical conditions. The evaluation metrics employed for comparison included Intersection over Union (IoU), pixel accuracy, and mean accuracy, amongst others.

3.2. Results

We base the mmsegmentation [9] to implement CPSeg. The results of our experiment were encouraging and provided substantial evidence in favor of our proposed method. From a quantitative perspective, our method consistently outperformed the state-of-the-art segmentation methods on all the evaluation metrics. For instance, the average IoU score for our method was significantly higher than that of other methods as shown in Figure 4, and the segmented images showed that our method was able to accurately segment the various regions in the flood images, such as water bodies, vegetation, and urban areas.

A detailed qualitative analysis, as presented in Table 1, further attests to the efficacy of our proposed methodology. DeepLab V3+ and PSPNet primarily use image-based segmentation, whereas DenseCLIP combines standard CLIP prompts with image segmentation masks. Our observations underscore the value of incorporating textual descrip-

Method	Building Flooded	Building Non-Flooded	Road Flooded	Road Non-Flooded	Water	Tree	Vehicle	Pool	Grass	mIoU
ENet [21]	6.94	47.35	12.49	48.43	48.95	68.36	32.26	42.49	76.23	42.61
DeepLabV3+ [8]	32.7	72.8	52.00	70.2	75.2	77.00	42.5	47.1	84.3	61.53
SegFormer-B0 [30]	70.81	79.04	69.09	85.27	80.86	86.06	56.02	66.13	91.06	76.20
PSPNet [38]	68.93	89.75	82.16	91.18	92.00	89.55	46.15	64.19	93.29	79.69
DenseCLIP [24]	72.98	79.55	65.94	84.48	78.97	85.74	55.01	63.74	90.92	75.14
CPSeg	**75.54**	**92.12**	**83.46**	**91.24**	**92.01**	**93.21**	48.01	64.15	**94.21**	**82.43**

Table 1. Per-class results with IOU and mIoU on FloodPrompt testing set.

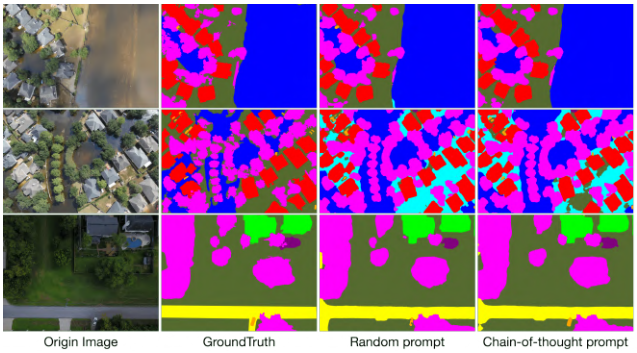

Figure 5. The performance with random prompting vs chain-of-thought prompting.

tions into segmentation, leading to richer context-aware outcomes. For example, when a text description highlighted a flooded street, our method adeptly identified and segmented this specific region in the image, a task at which traditional segmentation methods often faltered. This success highlights the utility and impact of the chain-of-thought process and language-guided context in enhancing image segmentation performance.

In conclusion, our experimental results unequivocally demonstrated the superiority of our proposed chain-of-thought, language-guided context-aware segmentation method over traditional image segmentation methods. By leveraging the power of language data, we were able to enhance the performance of image segmentation, particularly in the challenging and relatively unexplored domain of remote sensing imagery in flood scenarios.

4. Ablation Study

Firstly, we evaluated the impact of the chain-of-thought process. To do this, we compared the performance of our full model with a version that excluded the chain-of-thought process. We experiment different prompts, which standard prompt, random prompt, two prompts, and chain-of-thought prompt, with the mIoU. The objective was to quantify the effect of sequentially injecting textual information into the image segmentation process. Our results in Table 2 showed a substantial decrease in segmentation performance when the chain-of-thought process was omitted. The associated standard and random prompts provide only a macro-level cue, with the latter randomly selecting one of the available prompts.

Prompts	mIoU ↑
Standard prompt	75.26
Two prompts	75.98
Random prompt	76.23
Chain-of-thought prompt	82.43

Table 2. Ablation study with various prompting learning.

Specifically, the IoU score is higher indicating that the chain-of-thought process indeed plays a vital role in improving segmentation precision. This affirmed our hypothesis that a continual information stream could enhance the understanding of the scene, leading to improved segmentation outcomes.

Our FloodPrompt data is finer-grained classes segmentation tasks, we also set experiments for our network for the finer-grained semantic segmentation. We combine the non-flooded building and flooded building for a building class and non-flooded road and flooded road for the road classes. Table 4 indicate if we combine the flooded and non-flooded building and road for the same classes, building, and road. Our CPSeg still work for the combined labels with the chain-of-thought prompting with implicit learning.

In addition to performance comparisons, we analyzed the computational efficiency of our method relative to the baseline DenseCLIP, which also employs a Vision-Language model. Experiments were conducted on an NVIDIA A100 GPU, processing images of size 1024×1024. Our model, CPSeg, exhibits lower computational complexity in terms of both parameters and floating point operations per second (FLOPs), as shown in the Table 5. The recorded inference time for CPSeg is 42.85 FPS. These findings underscore the advantage of our chain-of-thought approach in handling finer-grained semantic segmentation

Object Class	Images (Flooded/Non-Flooded)	Images (Total)	Instances (Total)
Building	275/1272	(275+1272)	(3573+5373)
Road	335/1725	(335+1725)	(649+3135)
Vehicle	-/1105	1105	6058
Pool	-/676	676	1421
Tree	-/2507	2507	25889
Water	-/1262	1262	1784

Table 3. For experimental purposes in our CPSeg pipeline, we have compared the number of finer-grained segmentation classes and consolidated number within the same class.

Data Type	mIoU ↑
Original Data	82.43
Combined Data	87.89

Table 4. Segmentation analysis with different type data.

Method	FLOPs(G)	Params(G)	Inf time (fps)
DenseCLIP	1043.1	105.3	44.56
CPSeg	1037.4	100.8	42.85

Table 5. Performance with baseline for segmentation analysis.

tasks, offering not only superior performance but also increased efficiency compared to existing methods.

In conclusion, our ablation study provided valuable insights into the functioning of our proposed method. The results clearly showed that both the chain-of-thought process and the text encoder are crucial for our method's superior performance, thus justifying their inclusion in the framework. Furthermore, our study served to emphasize the significance of a detailed component-wise analysis in understanding and refining complex methodologies.

5. Discussion

While our research yields promising results, it is critical to recognize its constraints. Our experiments are predicated on a specialized dataset, targeting a specific disaster scenario - floods. Therefore, the efficacy of the proposed CPSeg method might differ with varying dataset and disaster contexts, warranting exploration in diverse scenarios like forest fires or earthquakes. Additionally, our current chain-of-thought process predominantly relies on textual cues. Hence, integrating other forms of context-sensitive data, including spatial or temporal aspects, might enhance the model's performance.

Strategically mapping these components, from an overarching view down to nuanced intricacies, is of paramount importance in this domain. As we progress, the pursuit of integrating multi-modal facets via the 'chain of thought' framework holds substantial promise. Refinement of the chain-of-thought process could potentially yield significant improvements in segmentation performance. Further, the incorporation of diverse forms of contextual data could augment the versatility and robustness of our framework. There also lies the intriguing possibility of exploring zero-shot learning in conjunction with chain-of-thought prompts from pre-trained Vision-and-Language models. Contrary to the conventional CLIP-based methodologies for image segmentation, our approach underscores the profound learning of image representation, which is achieved through the systematic provision of progressively granular prompts. Lastly, investigating the application of CPSeg other various scenarios could not only validate its effectiveness across contexts but also highlight areas requiring further improvement.

6. Conclusion

In conclusion, this study introduces CPSeg, a pioneering approach that employs a novel "Chain-of-Thought" process, leveraging language prompting to achieve finer-grained semantic segmentation in the field of remote sensing imagery. The work target particular applicability in flood disaster scenarios, capitalizes on the textual descriptions associated with images to enhance semantic understanding and improve segmentation performance. Through comprehensive validation, CPSeg demonstrates remarkable efficacy, pushing the boundaries of large-scale, context-aware data utilization and opening up new avenues for advancements in this domain. The results of this study offer valuable insights and contribute to the growing body of research on vision-language integration and its applications in remote sensing. The generalized methodology is amenable to application in alternative image scenarios, provided that comprehensive contextual details are available. We believe the approach based on chain-of-thought ideas can greatly enhance high-level tasks such as segmentation, detection, and regression.

Acknowledgments

This work was supported by the DeepCrop project and PerformLCA project (UCPH Strategic plan 2023 Data+ Pool).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 2

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2

[3] Martin Brandt, Compton J. Tucker, Ankit Kariryaa, Kjeld Rasmussen, Christin Abel, Jennifer Small an d Jerome Chave, Laura Vang Rasmussen, Pierre Hiernaux, Abdoul Aziz Diouf, Laurent Kergoat, Ole Mertz, Christian Ig el, Fabian Gieseke, Johannes Schöning, Sizhuo Li, Katherine Melocik, Jesse Meyer, Scott Sinno, Eric Romero, Erin Glennie, Amandine Montagu, Morgane Dendoncker, and Rasmus Fensholt. An unexpectedly large count of trees in the western Sahara and Sahel. *Nature*, 587:78–82, 2020. 1

[4] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 3

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 3

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 7

[9] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 6

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[11] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 540–557. Springer, 2022. 2

[12] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. 1, 2

[13] Lei Li. Edge aware learning for 3d point cloud, 2023. 1

[14] Lei Li. Segment any building for remote sensing, 2023. 1

[15] Lei Li, Tianfang Zhang, Stefan Oehmcke, Fabian Gieseke, and Christian Igel. Buildseg buildseg: A general framework for the segmentation of buildings. *Nordic Machine Intelligence*, 2(3), 2022. 1

[16] Lei Li, Tianfang Zhang, Stefan Oehmcke, Fabian Gieseke, and Christian Igel. Mask-fpan: Semi-supervised face parsing in the wild with de-occlusion and uv gan. *arXiv preprint arXiv:2212.09098*, 2022. 1

[17] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3

[18] Quande Liu, Youpeng Wen, Jianhua Han, Chunjing Xu, Hang Xu, and Xiaodan Liang. Open-world semantic segmentation via contrasting and clustering vision-language embedding. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XX*, pages 275–292. Springer, 2022. 3

[19] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 3

[20] Stefan Oehmcke, Lei Li, Jaime C Revenga, Thomas Nord-Larsen, Katerina Trepekli, Fabian Gieseke, and Christian Igel. Deep learning based 3d point cloud regression for estimating forest biomass. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, pages 1–4, 2022. 1

[21] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*, 2016. 7

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[23] Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and Robin Roberson Murphy. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9:89644–89654, 2021. 1, 2, 6

[24] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022. 1, 3, 7

[25] Jaime C Revenga, Katerina Trepekli, Stefan Oehmcke, Rasmus Jensen, Lei Li, Christian Igel, Fabian Cristian Gieseke,

and Thomas Friborg. Above-ground biomass prediction for croplands at a sub-meter resolution using uav–lidar and machine learning methods. *Remote Sensing*, 14(16):3912, 2022. 1

[26] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 3

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[28] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021. 3

[29] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022. 2

[30] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 1, 7

[31] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 2

[32] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 736–753. Springer, 2022. 1, 2

[33] Sukmin Yun, Seong Hyeon Park, Paul Hongsuck Seo, and Jinwoo Shin. Ifseg: Image-free semantic segmentation via vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2967–2977, 2023. 1

[34] Tianfang Zhang, Lei Li, Siying Cao, Tian Pu, and Zhenming Peng. Attention-guided pyramid context networks for detecting infrared small target under complex background. *IEEE Transactions on Aerospace and Electronic Systems*, 2023. 1

[35] Tianfang Zhang, Lei Li, Christian Igel, Stefan Oehmcke, Fabian Gieseke, and Zhenming Peng. Lr-csnet: Low-rank deep unfolding network for image compressive sensing. In *2022 IEEE 8th International Conference on Computer and Communications (ICCC)*, pages 1951–1957, 2022. 1

[36] Yicheng Zhang, Lei Li, Li Song, Rong Xie, and Wenjun Zhang. Fact: fused attention for clothing transfer with generative adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12894–12901, 2020. 1

[37] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 3

[38] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *Proceedings of the European conference on computer vision (ECCV)*, pages 267–283, 2018. 7

[39] Chong Zhou, Chen Change Loy, and Bo Dai. Denseclip: Extract free dense labels from clip. *arXiv preprint arXiv:2112.01071*, 2021. 3, 6

[40] Changsheng Zhou, Chao Yuan, Hongxin Wang, Lei Li, Stefan Oehmcke, Junmin Liu, and Jigen Peng. Multi-scale pseudo labeling for unsupervised deep edge detection. *Knowledge-Based Systems*, page 111057, 2023. 1