TEXT AS A RICHER SOURCE OF SUPERVISION IN SEMANTIC SEGMENTATION TASKS

Valerie Zermatten[1], Javiera Castillo Navarro[1], Lloyd Hughes[1], Tobias Kellenberger[2], Devis Tuia[1] *

[1] Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

[2] Federal Office of Topography swisstopo, Switzerland.

January 2023

Abstract

This paper introduces TACOSS a text-image alignment approach that allows explainable land cover semantic segmentation by directly integrating semantic concepts encoded from texts. TACOSS combines convolutional neural networks for visual feature extraction with semantic embeddings provided by a language model. By leveraging contrastive learning approaches, we learn an alignment between the visual and the (fixed) textual representations. In addition to producing standard semantic segmentation outputs, our model enables interactive queries with RS images using natural language prompts. The experimental results obtained on 50cm resolution aerial data from Switzerland show that TACOSS performs similarly to a standard semantic segmentation model while allowing the flexible usage of in- and out-of-vocabulary terms for the interactions with the image.

1 Introduction

Traditional methods for semantic segmentation associate a learned set of spectral and spatial characteristics with a target class. Classes are usually learned as exclusive, ignoring the thematic similarities between them: models penalize similarly a naive mistake e.g. classifying a river as a lake, as an aberrant prediction e.g. classifying the river as a road. A way to enforce this kind of semantic awareness is to use contextual information about the classes. If the remote sensing literature has widely explored class co-occurrence in space [1],

*Thanks to the Federal Office of Topography for funding.

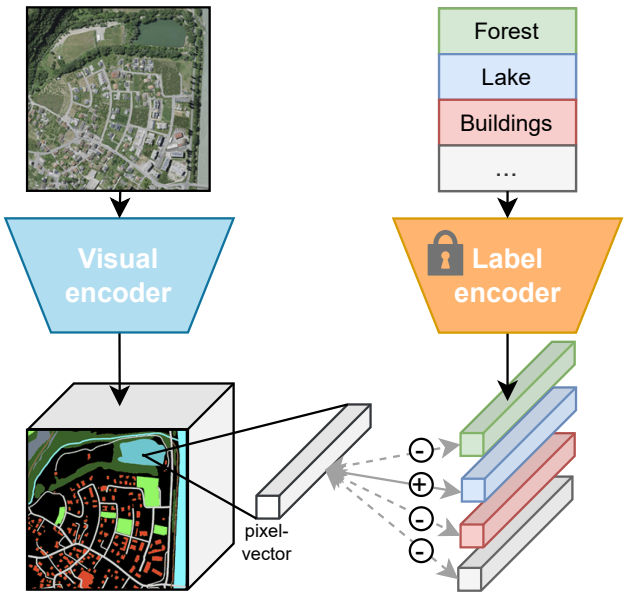

Figure 1: Overview of TACOSS: the visual features are extracted by the visual encoder and projected to a high-dimensional vector. A contrastive loss pushes the visual representation of each pixel close to their true label embeddings (extracted by language model), in this example representing the word '*Lake*', whereas the embeddings of a large set of negative labels ('*Forest*', '*Building*', etc.) are used as negative representations.

new advances in vision-language models open opportunities for the exploration of class similarities in the semantics of their definitions. Approaches that employ word embeddings learned from large text corpora can capture the thematic relationships between words, such that similar concepts are mapped to nearby points in the latent space and thus reduce the risk of aberrant mistakes. Several works [2, 3, 4, 5] have demonstrated the potential of incorporating thematic information from complex text captions. Diversified labels force the model to learn fine-grained differences between input images and also account for the sometimes fuzzy boundaries between classes. These frameworks lead to more robust predictions, i.e. more reasonable mistakes, and demonstrated impressive performances for adapting to new tasks through zero-shot or few-shot learning.

In this work, we re-frame the semantic segmentation problem for a multi-modal setting between images and texts and define it as an alignment problem between the respective representations. We design a semantic segmentation network where pixels are represented by interpretable embeddings and learned by taking into account both traditional semantic segmentation targets and vision-language alignment objectives. Specifically, we use a contrastive objective that

forces the model to respect the thematic similarity between classes by pushing away different concepts and bringing together similar ones, as they are represented in the word embedding space. To present a richer vocabulary and to have a more diffuse latent representation, text labels are augmented using words with similar meanings in the embedding space. As a result, in addition to the standard image segmentation results, our method enables pixel-level interactions between text and images, enabling the mapping of words beyond the vocabulary used for learning the land cover categories. This opens opportunities for the exploration of new thematic classes (out-of-domain mapping), logical reasoning and simple transfer from one map nomenclature to the other.

2 Related work

Convolutional neural networks, and more recently vision transformers are considered state-of-the-art classifiers for land cover segmentation [6, 7]. Several works have explored the integration of semantic relationships between classes, enabling the model to learn patterns from different classes and enhance the general classification performance [1]. An explicit semantic similarity between classes can be introduced by using label embedding methods [2]. These approaches embed the target labels into a high-dimensional latent space, as opposed to using the one-hot encoding of the classes. Another way to embed semantic information is to use text as source of information. For instance, GloVe [8] captures the semantic and syntactic relationships between words, such that semantically similar concepts are mapped to nearby points in the latent space. More recently, new models emerged combining natural language processing and computer vision to enable users to engage with images using textual input. For instance, CLIP [3] is trained with a contrastive learning strategy on a large-scale dataset of image-text pairs. It learns to align various image captions with visual features at the image level. Similarly, several models worked such as DenseCLIP [4], GroupViT [5] or VLT [9] focused to align features at the pixel level for dense prediction tasks through a contrastive learning approach. However, models trained on natural images tend to have limited transfer capacity to RS images [3, 10].

More generally, our work is related to recent efforts in the usage of language to interact with remote sensing (RS) images [11]. For instance, researchers have explored the use of text-based queries for image captioning [3], image retrieval [12], or question-answering [13]. While these works tend to interact with the visual features at the image level, others learn representations at the pixel level. For instance, by leveraging the Segment-Anything Model [7], the recent Text2Seg [10] model is able to do instance segmentation on RS images based on textual prompts. We aim to go further in that direction by developing a method that allows direct interactions with RS images at the pixel level through any text prompt, to achieve a model that can be queried on demand with new concepts represented as text embeddings.

3 Methodology

Our Text-As-supervision-for-COntrastive-Semantic-Segmentation (TACOSS) network is shown in Figure 1. It is a semantic segmentation pipeline where the model learns to align visual features to a semantic vector embedding produced by a pre-trained language model. Thanks to this alignment, TACOSS predicts for each pixel a vector close to the word embedding of the corresponding land cover label. Since language models are trained to respect semantic similarities between concepts, we can technically use any text prompt in natural language and calculate its similarity to the visual embedding, therefore assessing how much new concepts 'react' to the visual information from the RS image.

Semantic features. We generate semantic features from the class names using two word-embeddings :

- **GloVe** [8] maps (groups of) words into a 300 dimensional space accounting for semantic similarity found in textual corpora. GloVe not only captures the syntactic relationships between words but also semantic and analogic relationships, i.e. words with similar meanings or used in similar contexts tend to have similar vector representations in the embedding space.

- We also use the pre-trained **CLIP text encoder** [3] to embed the land cover labels. This transformer-based text encoder was trained with a contrastive learning strategy on a large-scale dataset of image-text pairs. For each class, a 512 dimensional vector is returned which is the average hidden state of the words defining the land cover category.

Since the embeddings are limited to a small number of classes (see next section), we enlarged the number of target classes using a set of synonym words close to the class names. This has the double role of avoiding the solution to collapse to a single point per class and also to add diversity in the label embeddings. We used both hypernyms, i.e. words that encompass a broader category (house→buildings), or hyponyms, i.e. words that are more specific (forest→oak tree). The words used to describe a class are not exclusive, i.e. a word used to describe a class can also be present in a different but related class.

Visual features. The visual features are extracted from very high-resolution images using a DeepLabv3 model [14] with a ResNet-50 [15] backbone. In addition to the original network classifier, a second classifier is added to project the Atrous Spatial Pyramid Pooling (ASPP) outputs to a high-dimensional vector for each pixel.

Loss. Drawing inspiration from contrastive learning, we aim here to align the visual features with the semantic features. We do so by maximizing the similarity between positive pairs of features (pixel and text refer to the same class) and minimizing the similarity between negative pairs (a pixel is compared to the embedding of the name of another class). Only the text embeddings of the original 16 classes are considered as positive representations, whereas the set of all synonyms are considered as negative representations. Since contrastive frameworks are known to benefit from abundant and diverse negative samples,

20 synonym words are searched for each class label. Following [16], we randomly sample $k = 2048$ pixels per batch and compute the contrastive loss as the InfoNCE loss [17]:

$$\mathcal{L}_{con} = -\sum_{i \in [0,k]} \log \frac{\exp\left(cos(z_i, z_p)/\tau\right)}{\sum_{a \in A(i)} \exp\left(cos(z_i, z_a)/\tau\right)} \quad (1)$$

with *cos* the cosine similarity between two vectors, z_i the network representation learned for pixel i, z_p is the positive semantic embedding of its ground truth class label p. z_a is the negative semantic embedding from the set containing all the negative land cover labels $A(i)$.

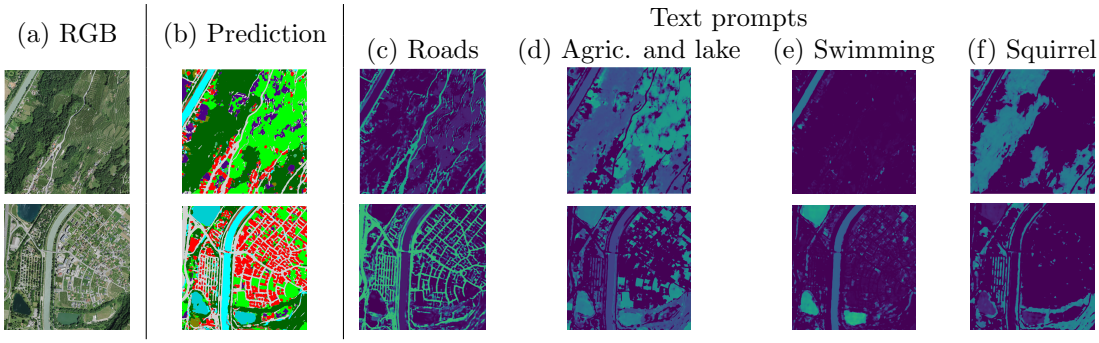

Figure 2: Illustration of the input images (a), semantic segmentation outputs (b) and interactions with different text prompts from the TACOSS-GloVe model: (c) 'Roads', (d) Agriculture and lake', (e)'Swimming' (f)'Squirrel'. The interaction outputs are normalized on a scale from 0 to 1, with dark blue color for low value and light green for higher values. Labels for semantic segmentation maps (among 16 classes): river, agricultural areas, buildings, forest, lake, roads, vineyards

4 Data and experimental set-up

Setup. The experiments below compare a baseline model trained with a cross-entropy loss, to our proposed TACOSS method with the two word embedding methods. The TACOSS visual backbone is trained for 300 epochs with a standard cross-entropy approach on one-hot encoded labels with the Adam [18] optimizer and a learning rate of $5 * 10^{-4}$with polynomial scheduling. Then the contrastive classifier is added and the model is further trained for 300 epochs with the contrastive objective. All models are trained with a batch of size 16 on input images of size 500×500, with colour augmentation, vertical and horizontal flips and random crops of size 200. The temperature parameter for the contrastive loss τ is empirically fixed to 0.03. The best hyper-parameters are fine-tuned independently on the validation set. To obtain the predicted labels

Table 1: Quantitative comparison of semantic segmentation performance between the baseline approach using a cross-entropy loss and our proposed TACOSS method with both CLIP and GloVe encoder.

	OA	**mIoU**	**mF1**
Baseline	57.04	25.28	35.78
TACOSS-GloVe	**57.18**	**25.85**	**36.46**
TACOSS-CLIP	55.61	24.82	35.81

from the model output, we compute a dot product between the model outputs and the semantic vectors representing the categories. Each pixel is attributed to the land cover category with the highest similarity.

Data. The model is trained on a real-world dataset provided by the Swiss Federal Office of Topography (swisstopo) on a study area located in Southwest Switzerland. The aerial images with RGB bands cover $63km^2$ with $50cm$ resolution. The pixel-level labels were produced by swisstopo annotators for the Swiss topographic landscape model (TLM) and span 16 different land cover classes including alpine, agricultural and urban land cover categories.

5 Results

As we can see in Table 1, the TACOSS pipeline using GloVe embeddings performs better on mean intersection over union (mIoU), macro F1-score (mF1) and overall accuracy (OA). A similar approach using CLIP embeddings remains close to the baseline results but does not outperform it, except for the mF1. We hypothesize that this is due to the larger number of dimensions that need to be aligned for CLIP embeddings (512) than for GloVe (300).

Figure 2 (b) illustrates the semantic segmentation maps generated by our model. The models identify correctly the main land cover classes such as water areas, buildings, forest and agricultural lands. However, the limits between categories tend to be blurry.

Figure 2 (c-f) presents some illustrations of interactions between TACOSS outputs with some natural language text prompts. The interaction map is obtained by computing the dot product between the output of TACOSS and the word embeddings of the text prompt. TACOSS is able to accurately segment new text prompts that correspond to land cover labels (in-vocabulary prompts, see 2 (c-d): these could be, for instance, alternative map nomenclatures used by other mapping agencies. But more interestingly, our method is able to go beyond the fixed set of labels and combine two land cover classes in the interaction map (see 2 (e)), or interact with words that are out-of-vocabulary (see 2 (f-h) and not related to land cover. The observed similarities between visual features and text prompt can be attributed to the semantic similarity with the words employed for training, i.e. the *'squirrel'* prompt highlights pixels most similar to forest, which is encoded through the usage of word embeddings.

6 Conclusion and Future work

In this paper, we propose TACOSS, a method to align visual features with semantic concepts at the pixel level for very high-resolution remote sensing images. TACOSS performs on par with a standard semantic segmentation model, while allowing interactions at the pixel level with any text prompt. Our preliminary results suggest that GloVe embeddings rather than CLIP ones allow a better semantic encoding between the visual and semantic features. This study has provided valuable insights for future research. One possible direction is the incorporation of a larger and more diverse vocabulary for the set of labels to be able to recognise land cover features with increased precision.

References

[1] D. Tuia, M. Volpi, and G. Moser. Decision fusion with multiple spatial supports by conditional random fields. *IEEE Trans. Geosci. Remote Sens.*, 56(6):3277–3289, 2018.

[2] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *NeurIPS*, 2013.

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[4] 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 *CVPR*, 2022.

[5] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. GroupViT: Semantic segmentation emerges from text supervision. In *CVPR*, 2022.

[6] Linus Scheibenreif, Joelle Hanna, Michael Mommert, and Damian Borth. Self-supervised vision transformers for land-cover segmentation and classification. In *CVPR*, 2022.

[7] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[8] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *EMNLP*, 2014.

[9] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *ICCV*, 2021.

[10] Jielu Zhang, Zhongliang Zhou, Gengchen Mai, Lan Mu, Mengxuan Hu, and Sheng Li. Text2Seg: Remote sensing image semantic segmentation via text-guided visual foundation models. *arXiv preprint arXiv:2304.10597*, 2023.

[11] Devis Tuia, Ribana Roscher, Jan Dirk Wegner, Nathan Jacobs, Xiao Xiang Zhu, and Gustau Camps-Valls. Towards a collective agenda on AI for earth science data analysis. *IEEE Geoscience and Remote Sensing Magazine*, 9(2):88–104, 2021.

[12] Li Mi, Siran Li, Christel Chappuis, and Devis Tuia. Knowledge-aware cross-modal text-image retrieval for remote sensing images. *CDCEO*, 2022.

[13] Christel Chappuis, Valérie Zermatten, Sylvain Lobry, Bertrand Le Saux, and Devis Tuia. Prompt-RSVQA: Prompting visual context to a language model for remote sensing visual question answering. In *CVPR*, 2022.

[14] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[16] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *NeurIPS*, 2020.

[17] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.