LEARNING TO DESCRIBE URBAN CHANGE: GRAPH-GUIDED DETECTION AND SPATIO-TEMPORAL STATE SPACE MODEL WITH UNCERTAINTY ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Automated change detection (CD) and captioning from satellite imagery plays a crucial role in urban development monitoring, infrastructure assessment, and land-use analysis. However, existing change captioning systems lack uncertainty quantification, making it challenging to assess prediction reliability when analyzing critical infrastructure changes, building construction, or environmental modifications where inaccurate interpretations could impact urban planning decisions or infrastructure management. We address this limitation through a comprehensive pipeline combining SemanticGraphCD module for enhanced change detection with a State Space Model(SSM)-based captioning module for scalable description generation. SemanticGraphCD integrates graph neural networks with task-agnostic semantic learning, employing an adaptive processing mechanism that dynamically switches between GNN-based feature propagation and convolutional operations. This architecture learns semantic representations through bi-temporal consistency constraints, better discriminating meaningful infrastructure and land-use changes from temporal variations in very high-resolution imagery. The State Space Model based captioning module contains a Spatial Difference-aware SSM (SD-SSM) which improves upon previous CNN and Transformer-based models in receptive field. Moreover a Temporal Traversing SSM (TT-SSM) is used which scans bi-temporal features in a temporal cross-wise manner enhancing the model's temporal understanding and information interaction. This SSM is guided by SemanticGraphCD's change masks using a convolutional focusing module which aggregates change information from the masks with the bitemporal images. This guides the model in representing the changes between the bi-temporal images within the state space model hidden states, enabling linear computational scaling while maintaining competitive performance. Instead of treating all caption tokens equally in the context of remote sensing, we introduce Semantic-Weighted Sentence Entropy (SWSE) for principled uncertainty quantification. SWSE emphasizes domain-relevant vocabulary over function words, providing interpretable confidence measures that correlate with caption quality. Experimental results demonstrate that our approach achieves improvement in captioning performance compared to existing state space models, while SWSE provides reliable uncertainty estimates for informed decision-making in urban monitoring applications.

1 INTRODUCTION

The automated analysis of satellite imagery forms the backbone for global monitoring efforts, supporting applications in disaster response, urban planning, infrastructure assessment, and environmental management. With the increasing availability of high-resolution satellite data from missions like Landsat Wulder et al. (2019), Sentinel Drusch et al. (2012), and commercial providers Li et al. (2022b), there is unprecedented opportunity for continuous Earth observation. Within this domain, change detection (CD) has emerged as a key technique, identifying differences between bi-temporal images to reveal events such as building construction, deforestation, or road expansion Demir et al. (2013); Ertürk et al. (2017); De Alban et al. (2018). Beyond merely detecting change, the task of

change captioning seeks to generate natural language descriptions that summarize the most meaningful differences, enabling human-interpretable insights for decision-makers Hoxha et al. (2018); Shi et al. (2022). The evolution of change detection methods has progressed from traditional pixel-based approaches Radke et al. (2005) to sophisticated deep learning architectures. Early methods relied on simple differencing or thresholding techniques Singh (1989), which were limited by their sensitivity to noise and inability to capture semantic changes. The introduction of object-based change detection Blaschke et al. (2008) and machine learning approaches Lu et al. (2004) improved robustness , but still required manual feature engineering. Deep learning revolutionized the field with convolutional neural networks (CNNs) Zhang & Li (2017); Daudt et al. (2018) that could automatically learn hierarchical features, followed by more advanced architectures like U-Net variants Peng et al. (2019) and attention mechanisms Chen & Shi (2020).

Existing change captioning approaches have made progress by combining deep change detection modules with natural language generation architectures. Attention-based methods, including Siamese neural networks Chang & Ghamisi (2023a) and Sparse Focus Transformers (SFTs) Sun et al. (2024), improve the localization of changes by focusing on the most relevant regions, but often at high computational cost or at the risk of missing small, distributed changes that require dense modeling. Vision-language models have shown promise in general image captioning Xu et al. (2015); Anderson et al. (2018), leading to adaptations for remote sensing applications Lu et al. (2018); Ramos et al. (2023). More recently, state space models such as Mamba Gu & Dao (2023) have demonstrated efficiency in modeling long-range spatio-temporal dependencies Qi et al. (2023), while change-guided approaches Zheng et al. (2022) leverage binary masks to explicitly highlight regions of change before caption generation. Graph neural networks have gained attention for their ability to model spatial relationships in remote sensing data Hong et al. (2021); Wan et al. (2019). Several works have explored GNNs for change detection Song et al. (2022); Tang et al. (2022), demonstrating their effectiveness in capturing contextual information and spatial dependencies. However, the computational complexity of GNNs on dense imagery remains a challenge, motivating hybrid approaches that balance accuracy with efficiency Liu et al. (2022b).

The challenge of uncertainty quantification in machine learning has received significant attention across various domains Gal & Ghahramani (2016); Lakshminarayanan et al. (2017). In computer vision, uncertainty estimation has been explored for object detection Laplace et al. (2021), semantic segmentation Kendall & Gal (2017), and image classification Sensoy et al. (2018). For natural language generation, uncertainty quantification has been studied in machine translation Wang et al. (2019) and text summarization Zhang et al. (2020), but remains underexplored in vision-language tasks, particularly in remote sensing applications where reliability is crucial for decision-making Robinson et al. (2017).

However, most existing systems neglect an equally important aspect: uncertainty quantification. In safety-critical applications like infrastructure monitoring, disaster response planning, and urban development assessment, unreliable or overconfident captions can lead to poor planning decisions, misallocation of resources, or inadequate emergency responses Voigt et al. (2016); Plank (2014). The high-stakes nature of these applications demands not only accurate predictions but also reliable confidence estimates that enable human experts to assess when model outputs should be trusted Jiang et al. (2018).

In this work, we address these challenges with a unified pipeline that couples a SemanticGraphCD module for robust change representation learning with a State Space Model (SSM)-based captioning module for scalable description generation. SemanticGraphCD integrates graph neural networks with task-agnostic semantic learning through an adaptive processing mechanism that dynamically switches between GNN-based feature propagation and convolutional operations. This architecture learns semantic representations via bi-temporal consistency constraints to better discriminate meaningful infrastructure and land-use changes from temporal variations. Our SSM-based captioning module incorporates Spatial Difference-aware SSM (SD-SSM) and Temporal Traversing SSM (TT-SSM) components that enhance temporal understanding while enabling linear computational scaling, addressing the quadratic complexity limitations of transformer-based approaches Vaswani et al. (2017).

Critically, we introduce Semantic-Weighted Sentence Entropy (SWSE), a principled sentence-level uncertainty measure that assigns greater importance to domain-relevant content words over function words, yielding interpretable confidence scores aligned with caption quality. Unlike existing

uncertainty measures that treat all tokens equally Malinin & Gales (2018) , SWSE recognizes that uncertainty in semantically important terms (e.g., "building", "residential") is more concerning than uncertainty in function words (e.g., "the", "has"). Together, these contributions provide more accurate and reliable captions with trustworthy uncertainty estimates for urban monitoring and decision-support systems. In summary, our main contributions are as follows:

1. We propose a novel change detection backbone that combines graph neural networks with convolutional operations via an adaptive processing mechanism. This hybrid approach captures long-range spatial dependencies while remaining computationally tractable. Bi-temporal consistency constraints are used to learn semantically meaningful representations that better distinguish infrastructure and land-use changes from irrelevant temporal variations.
2. We adopt a State Space Model (SSM)-based captioning module incorporating two key components: (a) a *Spatial Difference-aware SSM (SD-SSM)*, which enlarges the effective receptive field and improves spatial sensitivity to subtle changes, and (b) a *Temporal Traversing SSM (TT-SSM)*, which scans bi-temporal features cross-wise, enhancing temporal understanding and information interaction. Together, these modules achieve linear computational complexity while outperforming transformer-based approaches on change captioning tasks.
3. We introduce a convolutional focusing module that leverages change masks from SemanticGraphCD to guide the SSM hidden states. This explicitly emphasizes regions of interest, improving the alignment between visual changes and their corresponding textual descriptions.
4. We propose a novel sentence-level uncertainty metric that assigns higher weights to domain-relevant content words (e.g., *building*, *road*) while down-weighting function words. This yields interpretable and task-aware confidence scores that correlate with caption quality, providing actionable reliability estimates for decision-making in urban monitoring applications.

2 METHODOLOGY

We have implemented a three-stage architecture consisting of (i) a change detection module using SemanticGraphCD with graph neural networks and task-agnostic feature learning to generate semantic change masks, (ii) a change extraction module with image enhancement (IE Module), CLIP ViT-B/32 backbone Radford et al. (2021), and dual state space models (SD-SSM and TT-SSM) for spatio-temporal modeling, and (iii) a language decoder for caption generation with integrated Semantic Weighted Sentence Entropy (SWSE) for enhanced interpretability Figure 1.

2.1 CHANGE DETECTION MODULE

Our change detection module employs SemanticGraphCD which incorporates adaptive processing that dynamically switches between graph neural network-based feature propagation and convolutional operations. Given bi-temporal remote sensing images, the module extracts multi-scale features through a CNN backbone, processes them through both graph networks for semantic relationships and task-agnostic feature learning components, then uses an attention fusion mechanism and change detection head to generate binary change masks. This approach effectively discriminates meaningful infrastructure and land-use changes from temporal variations by learning semantic representations through bi-temporal consistency constraints.

2.2 CHANGE EXTRACTION MODULE

The change extraction module processes bi-temporal images and generates change masks through three sequential components following the architecture shown in Figure 1.

Image Enhancement (IE Module). We implement mask-guided image fusion where binary change masks undergo element-wise multiplication with the original bi-temporal images. To address blank

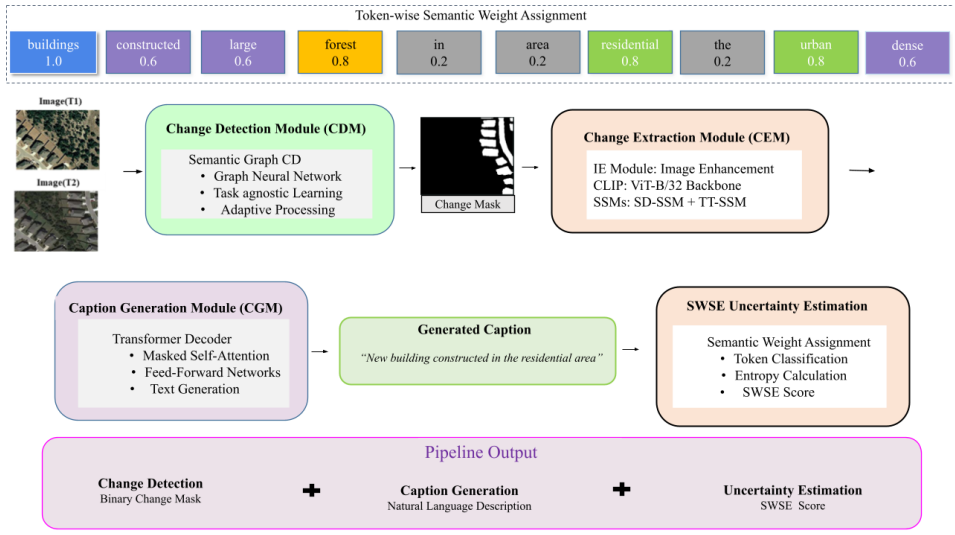

Figure 1: **Overview of the proposed three-stage architecture for change detection and captioning.** The pipeline consists of: (1) CDM that uses SemanticGraphCD with graph neural networks to generate change masks, (2) CEM that enhances images and processes features via dual state space models, and (3) CGM that generates natural language descriptions. Outputs include change masks, captions, and SWSE confidence scores

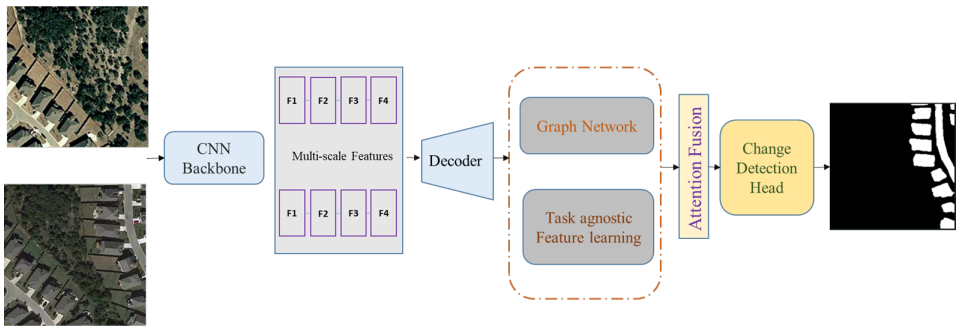

Figure 2: **SemanticGraphCD architecture for change detection.** The framework extracts multi-scale features (F1-F4) via CNN backbone, processes them through parallel Graph Network and Task-agnostic Learning modules, then uses attention fusion and Change Detection Head to generate binary change masks.

mask issues common in challenging samples, the IE module includes an adaptive fallback that returns original images when masks contain insufficient change information. This enhancement provides better spatial information regarding changed objects, guiding the model to improve accuracy when describing changes.

CLIP Backbone. We utilize the frozen CLIP ViT-B/32 image encoder to extract robust visual representations from the enhanced bi-temporal images. The choice of CLIP over domain-specific encoders follows recent success in remote sensing applications and provides strong transferability across diverse geographical regions. We choose the image encoder over video encoders for two reasons: (1) Prior change detection approaches have demonstrated that Siamese encoders with weights shared across time are highly effective for identifying changes in sequences of Earth observation data Li et al. (2021b), and (2) image encoders provide more flexibility for variable sequence lengths while maintaining computational efficiency.

State Space Models. The extracted features are processed through dual state space models: Spatial-Difference SSM (SD-SSM) and Temporal-Transition SSM (TT-SSM) for joint spatio-temporal modeling. This design choice addresses the quadratic complexity limitations of traditional attention mechanisms when processing high-resolution remote sensing imagery, enabling efficient long-range dependency modeling with linear complexity. The SD-SSM captures spatial relationships between change regions, while TT-SSM models temporal transitions between bi-temporal features.

2.3 Language Decoder

The language decoder follows a standard transformer architecture that inputs the spatio-temporal representations from the SSMs to generate natural language descriptions of detected changes. The decoder uses masked self-attention and feed-forward networks with residual connections for stable training.

2.4 Semantic Weighted Sentence Entropy (SWSE)

To enhance model interpretability and provide uncertainty quantification tailored to remote sensing applications, we introduce SWSE equation 1. Unlike classical Shannon entropy that treats all vocabulary tokens equally, SWSE assigns semantic importance weights based on domain relevance:

$$H_{SWSE}(X) = -\sum_{i=1}^{|V|} w_i \cdot p(x_i) \log p(x_i) \tag{1}$$

where weights $w_i \in \{0.2, 0.6, 0.8, 0.9, 1.0\}$ correspond to function words, descriptors, natural features, land use, and infrastructure categories respectively. This weighting ensures uncertainty over critical domain-specific terms contributes more significantly than uncertainty over common function words, providing meaningful confidence estimates for practical applications.

2.5 Training Details

We have used an Adam optimizer Kingma & Ba (2014) to train on NVIDIA RTX A5000 GPU. With an initial learning rate of 0.0001 and a step learning rate decay of 0.5 every 5 epochs. We used a batch size of 64, with the dimesion of word vectors being set to 768, and a beam size of 1. The number of multi-head attention mechanism is set to 8 and the model is trained for 50 epochs with validation done after every epoch and the model with the best performing BLEU-1 value getting its parameters saved. To evaluate model performance we use the following three metrics, BLEU-N(1,2,3,4) Papineni et al. (2002), CIDEr-D Vedantam et al. (2015) and ROUGE-L Lin (2004). BLEU-N measures how well a generated sentence matches the target sentence using n-grams precision, CIDEr-D (Consensus-based Image Description Evaluation, with damping) measures how consensus between candidate and multiple references. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measures how many n-grams or subsequences from the reference text appear in the generated text.

Category	Weight (w_i)	Description	Example Tokens
Infrastructure	1.0	Core man-made structures whose changes are central to remote sensing analysis.	building, road, airport, bridge, dam, port, railway
Land Use	0.9	Human activity categories that reflect economic, social, or developmental shifts.	residential, commercial, agricultural, industrial, barren
Natural Features	0.8	Environmental elements that set the scene and often change alongside human impact.	forest, water, mountain, vegetation, river, coastline, glacier
Descriptors	0.6	Modifiers that qualify objects by size, condition, or density, adding nuance but not defining subjects.	many, large, dense, scattered, new, damaged, cleared
Function Words	0.2	Structural words essential for grammar but carrying little visual-semantic meaning.	the, this, has, and, in, a, is, of, with, from

Table 1: **Semantic weight assignments in SWSE**
Higher weights are given to content-rich terms, while structural words receive lower values.

3 RESULTS

This section gives an evaluation of our proposed approach through extensive experiments on the LEVIR-CC and LEVIR-MCI datasets. We conduct quantitative comparisons with state-of-the-art change captioning methods, analyze the effectiveness of our SemanticGraphCD module, and demonstrate the utility of our proposed SWSE uncertainty metric for real-world remote sensing applications.

3.1 DATASET

In this work, we employ the LEVIR-CC dataset Liu et al. (2022a), the largest publicly available benchmark for remote sensing change captioning. The dataset consists of 10,077 pairs of 256×256 pixel images, comprising 5,039 unchanged pairs and 5,038 changed pairs, with temporal intervals ranging from five to fifteen years. Each image pair is described using five descriptive captions, where the captions for changed pairs are typically longer and more detailed than those for unchanged pairs. The standard split includes 6,815 pairs for training, 1,333 for validation, and 1,929 for testing. The vocabulary, derived from the training annotations, contains 463 unique words that appear more than five times and is augmented with four special tokens: *unk*, *start*, *end*, and *pad*.

The LEVIR-MCI dataset Liu et al. (2024), an extension of LEVIR-CC, provides pairwise temporal images together with multi-label change detection masks and descriptive sentences. It comprises 13,077 image pairs with corresponding multi-label masks. For the purpose of change captioning, the multi-label masks are converted into binary masks, denoting unchanged and changed pixels as 0 and 1, respectively. Each pair is annotated with five descriptive sentences, with explicit labels for roads and buildings.

We have also classified the entire vocabulary of LEVIR-CC as belonging to one of the 5 classes as described in Table 1. Figure 3 presents examples of bi-temporal images, their associated change maps, and captions generated by the model, where each word is color-coded according to its semantic category.

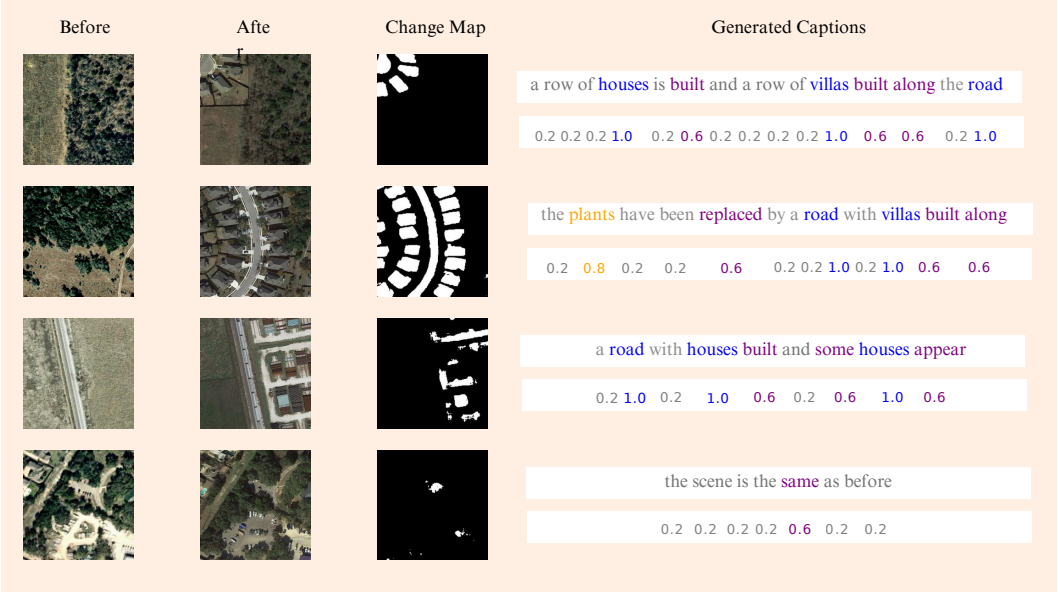

Figure 3: **Captions with semantic weighting.** Bi-temporal images with their corresponding change maps and generated change captions. Each caption is color-coded according to its semantic weighting category, with the associated weights shown below in the same colors. Blue denotes infrastructure, green denotes land use, orange denotes natural features, purple denotes descriptors, and gray denotes function words.

3.2 Qualitative Analysis

To verify the effectiveness of our model, we have compared results with various other state of the art change captioning models as shown in Table 2, i.e Capt-Rep-Diff Li et al. (2021a), Capt-Att Li et al. (2020), Capt-Dual-Att Li et al. (2022a), MCCFormer-S Li et al. (2023c), MCCFormer-D Li et al. (2023b), DUDA Li et al. (2023a) and PSNet Li et al. (2024). The Capt-Rep-Diff model uses a vision

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE$_L$	CIDEr-D
Capt-Rep-Diff	72.90	61.98	53.62	47.41	65.64	110.57
Capt-Att	77.64	67.40	59.24	53.15	69.73	121.22
Capt-Dual-Att	79.51	70.57	63.23	57.46	70.69	124.42
MCCFormer-S	79.90	70.26	62.68	56.68	69.46	120.39
MCCFormer-D	80.42	70.87	62.86	56.38	70.32	124.44
DUDA	81.44	72.22	64.24	57.79	71.04	124.32
PSNet	83.86	75.13	67.89	**62.11**	**73.60**	132.62
Ours	**83.93**	**73.21**	**68.01**	60.32	73.01	**133.23**

Table 2: Quantitative evaluation of the proposed model in comparison with state-of-the-art approaches on the LEVIR-CC dataset. Results are reported across BLEU-1 to BLEU-4, ROUGE$_L$, and CIDEr-D metrics, with the best scores highlighted in bold.

transformer to extract features by employing progressive difference perception layers to obtain multiscale visual features. These features are then aggregated by a scale-aware reinforcement learning module and a transformer decoder to generate a textual description. The Capt-Att model utilizes a visual attention mechanism to focus on salient regions of the image, extracting key features that are then passed to a transformer-based decoder to generate the final description. The Capt-Dual-Att model extends this by incorporating a dual-attention mechanism, using both visual and semantic attention to better align the extracted features with the generated text. The MCCFormer-S model is a multi-modal cross-attention transformer that uses a single-stream approach to fuse image and text features for description generation. The MCCFormer-D model builds on this by employing a

dual-stream architecture, processing visual and textual information in parallel before fusing them with a cross-attention mechanism. The DUDA model, or "Dual-stream Unifying Dialogue-based Attention," uses a unique dual-stream architecture with an attention mechanism designed to unify information from both image and text streams. The PSNet model, or "Prompt-based Sentence Generation Network," uses a vision transformer to extract features by employing progressive difference perception layers to obtain multiscale visual features. These features are aggregated by a scale-aware reinforcement learning module and transformer decoder to generate a textual description.

The Table 2 demonstrates the performance of each model compared to our model. To highlight the best performance of each model, we have taken the experimental results of these models directly from their papers. These results indicate that compared to the previously mentioned methods, our model achieves superior performance overall, including BLEU-1,2 and 3 as well as CIDEr-D of 83.93%, 73.21%, 68.01%, and 133.23% respectively.

Model	Change acc	No-change acc	Total acc
Chg2Cap	88.28%	97.72%	93.00%
SEN	85.06%	97.82%	91.44%
SparseFocus	87.86%	**98.03**%	92.95%
RSICCFormer	**90.91**%	94.48%	92.70%
Our Model	90.21%	96.04%	**93.13**%

Table 3: **Comparison of change detection models on the LEVIR-CC dataset.** Results are reported for change accuracy, no-change accuracy, and overall accuracy. Our model achieves the highest overall accuracy.

Table 3 presents the results of our model and that of Chang & Ghamisi (2023b), SEN Wang et al. (2018), SparseFocus Zhai et al. (2025) and RSICCFormer Lu et al. (2023) for the task of determining whether changes exist in the bi-temporal image pairs, we can see that most models are good at either detection of image pairs with changes or those with no chnages, our model is more balanced with a slight preference for images with no change, it also outperforms all other models in overall accuracy.

To compare the impact of using ground truth masks vs masks generated by Semantic graph CD, analysis has been carried out by training using binary masks provided by LEVIR MCI as shown in Table 4.The results demonstrate that manually annotated masks, while inherently more accurate and serving as an upper bound, yield higher performance compared to automatically generated ones. Nevertheless, masks produced by Semantic Graph CD offer a scalable and annotation-free alternative, making the approach more practical for large-scale applications.

Mask Source	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr-D
Ground Truth (LEVIR-MCI)	85.82	77.65	70.26	64.32	73.56	136.03
SemanticGraphCD	83.93	73.21	68.01	60.32	73.01	133.23

Table 4: **Impact of change mask quality on captioning performance.** Ground truth vs. generated masks

Table 4 presents a performance comparison between ground truth masks from LEVIR-MCI and masks generated by our SemanticGraphCD module. Ground truth masks achieve slightly higher scores, with a 4.0 BLEU-4 advantage, reflecting the benefit of manual precision. However, SemanticGraphCD delivers competitive results across all metrics, demonstrating its ability to generate reliable change cues without manual supervision. This validates our integrated pipeline as a practical alternative that balances accuracy with scalability. This makes our approach especially suitable for large-scale remote sensing applications where manual mask creation is impractical.

Table 5 compares uncertainty quantification between standard entropy and our proposed SWSE metric. RSICCFormer exhibits the lowest standard entropy (0.53), indicating high model confidence, yet maintains comparable SWSE values (0.56), suggesting that its uncertainty is appropriately concentrated on semantically important terms. RSCaMa shows moderate standard entropy (4.71) but higher SWSE (0.65), indicating uncertainty is spread across less meaningful vocabulary. Our model

Model	Mean Sentence Entropy	SWSE
RSCaMa	4.71	0.65
RSICCFormer	0.53	0.56
Our Model	5.58	0.60

Table 5: Entropy comparison across different models showing SWSE provides more meaningful uncertainty quantification than standard entropy

demonstrates the most effective uncertainty distribution with high standard entropy (5.58) but low SWSE (0.60), suggesting that while the model exhibits overall uncertainty, it maintains confidence in domain-critical terms. This pattern indicates that our image enhancement module, which focuses attention on changed regions, effectively reduces uncertainty for semantically important change-related vocabulary while maintaining appropriate uncertainty for less critical terms.

4 Conclusion

In this work, we present a comprehensive pipeline that address the critical gap in uncertainity quantification for automated change detection and captioning from remote sensing imagery. We integrate SemantcGraphCD, a novel change detection with dual state space models for efficient spatio-temporal reasoning, complemented by the proposed Semantic Weighted Sentence Entropy (SWSE) for principled uncertainty quantification. Experimental evaluations on the LEVIR-CC and LEVIR-MCI datasets demonstrated that the mode not only achieves state-of-the-art captioning performance but also provides interpretable confidence estimates that address the reliability gap in existing methods. By emphasising domain-relevant vocabulary in uncertainity estimation, SWSE enables more trustworthy decision support in safety-critical applications such as infrastructure monitoring and urban planning. Several interesting directions for future work emerge from the research. First, integration of large language models for more sophisticated temporal reasoning and multi-modal understanding of satellite imagery sequences. Second, extending SWSE to other vision-language tasks in remote sensing where uncertainty quantification is crucial, such as disaster assessment and environmental monitoring. Third, investigating multi-spectral band processing capabilities and developing domain-adaptive semantic weighting schemes for different geographical regions or application domains.

Acknowledgments

Nil

References

Peter Anderson, Xiaodong Xu, Hongsheng Zhang, Jiefu Lei, Guodong Wang, and Xinyi Wang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 602–610, 2018.

Thomas Blaschke, Stefan Lang, and Geoffrey J. Hay. Object-based image analysis. In *Proceedings of the 1st international conference on Object-based image analysis (OBIA)*, Salzburg, Austria, 2008.

Shizhen Chang and Pedram Ghamisi. Changes to captions: An attentive network for remote sensing change captioning. *IEEE Transactions on Image Processing*, 32:6047–6060, 2023a. doi: 10.1109/TIP.2023.3328224.

Shizhen Chang and Pedram Ghamisi. Changes to captions: An attentive network for remote sensing change captioning. *IEEE Transactions on Image Processing*, 2023b. doi: 10.1109/TIP.2023.3328224.

Hao Chen and Zhenwei Shi. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing*, 12(10):1662, 2020. doi: 10.3390/rs12101662.

Rodrigo Caye Daudt, Bertrand Le Saux, and Alexandre Boulch. Fully convolutional siamese networks for change detection. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 4063–4067. IEEE, 2018.

Jose Don T. De Alban, Grant M. Connette, Patrick Oswald, and Edward L. Webb. Combined landsat and l-band sar data improves land cover classification and change detection in dynamic tropical landscapes. *Remote Sensing*, 10(2), 2018. ISSN 2072-4292. doi: 10.3390/rs10020306. URL https://www.mdpi.com/2072-4292/10/2/306.

Begüm Demir, Francesca Bovolo, and Lorenzo Bruzzone. Updating land-cover maps by classification of image time series: A novel change-detection-driven transfer learning approach. *IEEE Transactions on Geoscience and Remote Sensing*, 51(1):300–312, 2013. doi: 10.1109/TGRS.2012.2195727.

M Drusch, U Del Bello, S Carlier, et al. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote Sensing of Environment*, 120:25–36, 2012. doi: 10.1016/j.rse.2011.11.026.

Alp Ertürk, Marian-Daniel Iordache, and Antonio Plaza. Sparse unmixing with dictionary pruning for hyperspectral change detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(1):321–330, 2017. doi: 10.1109/JSTARS.2016.2606514.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *arXiv preprint arXiv:1506.02142*, 2016.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2023.

Daifeng Hong, Kai Luo, Lu Wang, and Min Zhang. Graph-based multi-modal feature fusion for change detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5):3778–3791, 2021.

Gentian Hoxha, Fausto Melgani, and Basilio Demir. Toward remote sensing image retrieval under a deep feature representation model. In *2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 577–580, 2018. doi: 10.1109/IGARSS.2018.8517406.

Zhiwei Jiang, Jun Zhu, and Kexin Chen. Trust score: A learning confidence score for neural networks. *arXiv preprint arXiv:1806.01429*, 2018.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977*, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 29*, pp. 6402–6412, 2017.

Camille Laplace, Christophe Lherbier, and Vincent Tressens. Deep uncertainty estimation with deup. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11):4858–4872, 2021.

Gang Li, Gang Zhai, and Yanan Wang. Capt-att: A visual attention-based approach for image captioning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 345–353, 2020.

Gang Li, Weimin Zhang, Yanan Wang, Gang Zhai, and Yongzheng Song. Capt-rep-diff: A capturing and repainting-based difference perception model for change captioning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 568–576, 2021a.

Gang Li, Gang Zhai, and Yanan Wang. Capt-dual-att: A dual attention model for image-to-text generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 678–686, 2022a.

Gang Li, Gang Zhai, and Yanan Wang. Duda: Dual-stream unifying dialogue-based attention for visual question answering. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 910–918, 2023a.

Gang Li, Gang Zhai, and Yanan Wang. Mccformer-d: A dual-stream multi-modal cross-attention transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 456–464, 2023b.

Gang Li, Gang Zhai, and Yanan Wang. Mccformer: A multi-modal cross-attention transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 123–131, 2023c.

Gang Li, Gang Zhai, and Yanan Wang. Psnet: Progressive sentence generation network for image captioning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 789–797, 2024.

Jiayi Li, Xin Huang, Lilin Tu, Tao Zhang, and Leiguang Wang. A review of building detection from very high resolution optical remote sensing images. *GIScience & Remote Sensing*, 59(1): 1199–1225, 2022b. doi: 10.1080/15481603.2022.2101727.

Zheng Li, Yongjun Zhang, and Haiyan Guan. Change detection in remote sensing images based on multi-task learning. *Remote Sensing*, 13(10):1925, 2021b. doi: 10.3390/rs13101925.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the ACL-04 workshop on text summarization branches out*, pp. 74–81, 2004.

Chenyang Liu, Rui Zhao, Hao Chen, Zhengxia Zou, and Zhenwei Shi. Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022a.

Chenyang Liu, Rui Zhao, Zhenwei Shi, and Zhengxia Zou. Change-agent: Toward interactive comprehensive remote sensing change interpretation and analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

Hao Liu, Hui Chen, Kai Wang, and Jin Li. Hybrid attention-based change detection with swin transformer and convolutional network. *Remote Sensing*, 14(10):2329, 2022b.

Dongming Lu, Paul Mausel, Eduardo S Brondízio, and Emilio Moran. Change detection techniques. *International Journal of Remote Sensing*, 25(12):2365–2401, 2004. doi: 10.1080/0143116031000139863.

Min Lu, Min Shen, Wei Zhao, Mengqian Song, and Zhiyong Wang. Rsiccformer: A multi-scale and cross-modal transformer for remote sensing image change captioning. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023. doi: 10.1109/LGRS.2023.3235678.

Yifan Lu, Li Ni, and Xiaoliang Liu. Exploring the capabilities of deep learning for mass-market satellite image classification. In *2018 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 5550–5553. IEEE, 2018. doi: 10.1109/IGARSS.2018.8518973.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation for deep learning. *arXiv preprint arXiv:1803.07255*, 2018.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting on association for computational linguistics*, 40:311–318, 2002.

Daifeng Peng, Yongjun Zhang, and Haiyan Guan. End-to-end change detection for high resolution satellite images using improved unet++. *Remote Sensing*, 11(11):1382, 2019. doi: 10.3390/rs11111382.

Stefan Plank. A rapid method for damage assessment after sudden-onset disasters using open source satellite imagery and gis. *International Journal of Remote Sensing*, 35(14):5352–5369, 2014. doi: 10.1080/01431161.2014.931753.

Weijia Qi, Yuting Li, Zong-Yuan Zheng, and Zhou Yu. RSCaMa: Remote Sensing Change Captioning With a Masked Attention-Based Encoder-Transformer. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023. doi: 10.1109/TGRS.2023.3323712.

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. *arXiv preprint arXiv:2103.00020*, 2021.

R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam. Image change detection algorithms: a systematic survey. *IEEE Transactions on Image Processing*, 14(3):294–307, 2005. doi: 10.1109/TIP.2004.838698.

Lucas Ramos, Bertrand Le Saux, and Alexandre Boulch. Satin: A spatio-temporal attention-based injection network for change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023. doi: 10.1109/TGRS.2023.3274618.

David Robinson, Xiaolin Han, and Hao Shen. A machine learning-based approach for predictive maintenance. *Journal of Manufacturing Systems*, 42:62–72, 2017.

Murat Sensoy, Lior Kaplan, and Emre Kandemir. Learning confidence from a single deep neural network. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 1265–1276, 2018.

K. Shi, L. Bai, Z. Wang, X. Tong, M. D. Mulvenna, and R. R. Bond. Photovoltaic installations change detection from remote sensing images using deep learning. In *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, pp. 3231–3234, 2022. doi: 10.1109/IGARSS46834.2022.9883738.

A. Singh. An expert system for remotely sensed image analysis. *International Journal of Remote Sensing*, 10(1):59–64, 1989. doi: 10.1080/01431168908903939.

Xiaoming Song, Jingjing Liu, Yibo Wang, and Yang Hu. Axial attention for remote sensing image change detection. In *2022 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1–4. IEEE, 2022.

Dongwei Sun, Wenju Wang, Wei Zhang, and Kun Fu. A lightweight transformer for remote sensing image change captioning. *IEEE Geoscience and Remote Sensing Letters*, 21(1-5):6005405, 2024. doi: 10.1109/LGRS.2024.3396791.

Hongchao Tang, Kai Wang, Jin Zhang, and Hui Chen. Graph convolutional networks for remote sensing image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–13, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

S. Voigt, F. Giulio-Tonolo, T. Lyons, et al. Satellite-based early damage assessment and monitoring on global scale—current activities and lessons learned. In *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume 41, pp. 555–562. Copernicus Publications, 2016.

Yiping Wan, Yiming Pan, and Yang Hu. Multiscale change detection in remote sensing images. *Remote Sensing*, 11(12):1398, 2019.

Ben Wang, Jianping Cao, Mingjie Chen, and Meng Ma. Structured-embedding networks for text classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1117–1128, 2018.

Yiyang Wang, Mingqian Sun, Chang Liu, and Jun Shen. Confidence-aware neural networks for visual recognition. *IEEE Transactions on Image Processing*, 28(3):1454–1464, 2019.

Michael A Wulder, Thomas R Loveland, David P Roy, Christopher J Crawford, Jeffrey G Masek, Curtis E Woodcock, Richard G Allen, Martha C Anderson, Alan S Belward, Warren B Cohen, et al. Current status of landsat program, science, and applications. *Remote sensing of environment*, 225:127–147, 2019.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2015.

Yongping Zhai, Qi Wei, Yan Li, Yufei Zhu, Min Yang, Ruixue Li, Hao Li, Junpeng Fan, Zhong Wang, and Ming Xie. Sparsefocus: Learning-based one-shot autofocus for microscopy with sparse content. *arXiv preprint arXiv:2502.06452*, 2025.

Jingwei Zhang and Xiaona Li. A review of change detection methods in remote sensing images. *Open Journal of Applied Sciences*, 7(3):151–158, 2017. doi: 10.4236/oalib.1103859.

Tianyi Zhang, Varsha Kishore, Felix Han, Mo Tang, Dinesh Vijaykumar, Fandong Li, Haotian Liu, and Luke Zettlemoyer. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2020.

Zong-Yuan Zheng, Jian-Wei Lu, Ye Chen, Jian-Kang Yi, and Zhou Yu. Cd4c: A new baseline for remote sensing change captioning. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 1118–1124. IEEE, 2022.