Grounding Synthetic Data Generation With Vision and Language Models

Ümit Mert Çağlar, Alptekin Temizel
Graduate School of Informatics
METU, Turkey
mecaglar@metu.edu.tr, atemizel@metu.edu.tr

Abstract

Deep learning models benefit from increasing data diversity and volume, motivating synthetic data augmentation to improve existing datasets. However, existing evaluation metrics for synthetic data typically calculate latent feature similarity, which is difficult to interpret and does not always correlate with the contribution to downstream tasks.

We propose a vision-language grounded framework for interpretable synthetic data augmentation and evaluation in remote sensing. Our approach combines generative models, semantic segmentation and image captioning with vision and language models. Based on this framework, we introduce ARAS400k: A large-scale Remote sensing dataset Augmented with Synthetic data for segmentation and captioning, containing 100k real images and 300k synthetic images, each paired with segmentation maps and descriptions.

ARAS400k enables the automated evaluation of synthetic data by analyzing semantic composition, minimizing caption redundancy, and verifying cross-modal consistency between visual structures and language descriptions. Experimental results indicate that while models trained exclusively on synthetic data reach competitive performance levels, those trained with augmented data (a combination of real and synthetic images) consistently outperform real-data baselines. Consequently, this work establishes a scalable benchmark for remote sensing tasks, specifically in semantic segmentation and image captioning. The dataset is available at zenodo.org/records/18890661 and the code base at github.com/caglarmert/ARAS400k.

1. Introduction

Generative data augmentation is an effective alternative when collecting additional real data is costly or infeasible [8]. However, synthetic samples often lack explainability and interpretability [6]. Moreover, existing evaluation metrics often fail to capture the semantic alignment between synthetic and real images [2] and their measurement procedures are not transparent [4].

To address these limitations, we propose a three-stage workflow (Figure 1). In the first stage, we acquire and align true-color images with land cover data. After cleaning and pre-processing we produce an initial set of real image-segmentation map pairs. These pairs are then used to train semantic segmentation models, while a generative model is trained on the true-color images to generate synthetic samples. In the second stage, we utilize the segmentation models and create the segmentation maps of the synthetic samples. In the final stage we employ foundation models to generate descriptive captions by integrating visual content with composition statistics derived from the segmentation maps.

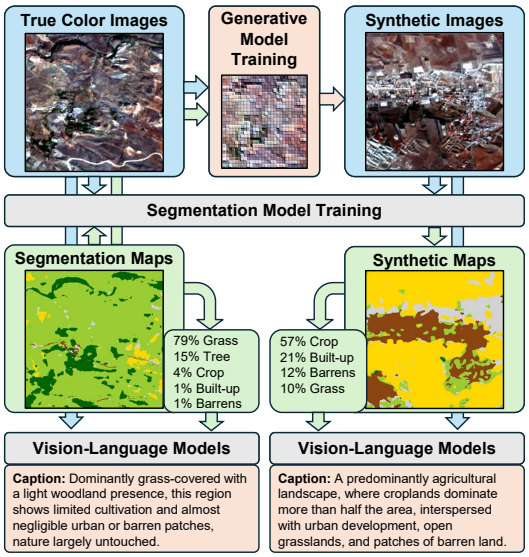

Figure 1. ARAS400k synthetic data augmentation framework.

The contributions of this work are summarized as follows:

- **Large-Scale Multi-Modal Dataset:** We introduce the fully open-sourced and publicly available ARAS400k, a comprehensive remote sensing dataset consisting of 100,240 real images and 300,000 synthetic images, each paired with semantic segmentation maps and over 2 million descriptive captions.

- **Automated Context-Aware Framework:** We propose an automated pipeline for context-aware caption generation and evaluation that utilizes composition statistics available from ground-truth or obtainable from segmentation models, reducing reliance on manual labor.
- **Vision-Language Integration:** We demonstrate a novel integration of foundation models to guide synthetic data evaluation through semantic consistency and redundancy reduction.
- **Proven Downstream Utility:** We provide an experimental benchmark showing that models trained on our augmented dataset (real + synthetic) consistently outperform those trained on real data alone, particularly in addressing class-imbalance for under-represented categories.

2. Related Works

Recent advances demonstrate that synthetic data can effectively substitute or augment real datasets. SynthCLIP [9] and SynGround [10] show that models trained exclusively on synthetic image-caption pairs can achieve performance comparable to real-data models in visual grounding and alignment tasks. The utility of such data is further enhanced by distillation techniques that prioritize high-quality samples over volume [26].

In the remote sensing domain, the need for precise spatial alignment and multi-scale feature integration is a challenge. Combining detail attention sampling with a teacher-student network effectively integrates local and global features, yielding state-of-the-art accuracy in high-resolution remote sensing image classification [18]. Research into multidimensional datasets suggests that the strategic sequencing of Generative Adversarial Networks (GANs) and feature selection is critical for maintaining model validity and generalization capability [1].

Although Denoising Diffusion Probabilistic Models [12] and newer variants have proven to be successful in generating images, they often require longer training and inference times compared to GAN architectures [28]. On the other hand, GAN models tend to have problems such as mode collapse, vanishing gradients, non converging or unstable training and they are sensitive to hyperparameters [3]. Various methodologies have been developed to address these limitations by improving convergence and output quality [13]. Recent advancements also explore hybrid approaches that combine the fast inference of GANs with the high fidelity of diffusion models through distillation techniques [14].

Finally, evaluating the large scale datasets and generative models requires more than standard metrics. The CLIPScore [11] metric is shown to align more with human assessment, enabling reference-free caption evaluation. Although it is especially crucial for non-referenced synthetic data, the ultimate validation for synthetic data lies in the ability to improve downstream tasks, such as semantic segmentation.

3. Methods

3.1. Data Acquisition

We acquired data from ESA Sentinel-2 RGBNIR true-color images and WorldCover 2021 [29] land cover maps. We then geo-aligned the land cover and true-color layers using their respective geographic coordinates and extracted 157,283 image patches at a resolution of 256×256 pixels.

The original land cover data contains 11 classes: tree, shrub, grass, crop, built-up, barren, snow, water, wetland, mangrove and moss. However, several classes are under-represented. Snow, wetland, moss and mangrove together appear only in 1,719 patches (1.1%). To address this imbalance, we merged snow into barren, mangrove and wetland into tree cover and moss into grass, resulting in seven final classes.

Patches dominated by water often had sun glare, which negatively affect image color and brightness, thus we removed 7,063 patches containing more than 90% water coverage. After filtering out these samples and removing 49,980 patches with missing values, we ended up with a total of 100,240 patches covering an area of $656{,}933\,\mathrm{km}^2$.

3.2. Data Generation Pipeline

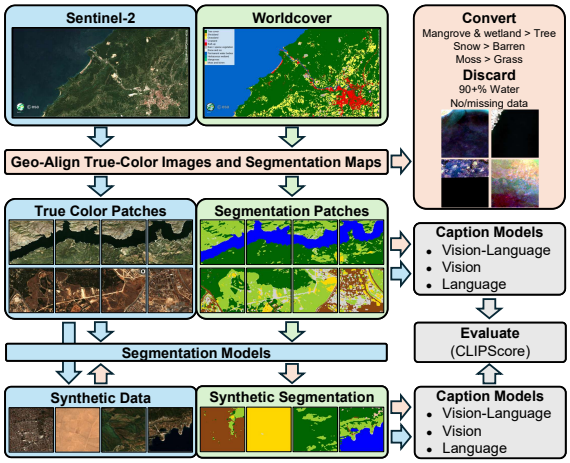

Figure 2. Data generation pipeline, comprised of data acquisition, alignment, cleaning, pre-processing, segmentation and generative model training, caption generation and evaluation.

We develop a cross-domain pipeline, illustrated in Figure 2, integrating visual and textual information to improve interpretability and contextual grounding. The pipeline performs generative data augmentation to increase data diversity and volume, and uses image captioning to translate visual content into language descriptions. This enables language models to evaluate both real and synthetic images using semantic and contextual information.

We updated the official implementation of the Alias-Free Generative Adversarial Networks (StyleGAN3) [15] to be

more efficient, utilizing the updated libraries and newer GPU architectures, enabling faster training and inference. With this implementation, we used real true-color images to train our StyleGAN3 model from scratch. We also combined the ideas from Semantic Image Synthesis with Spatially-Adaptive Normalization (SPADE) [20] and U-Net Based Discriminator for Generative Adversarial Networks [23] to train our custom segmentation map conditioned GAN architecture. We used real images and corresponding segmentation maps to generate synthetic augmentations of existing image-segmentation pairs.

Our framework uses fine-tuned, trained-from-scratch or pre-trained models for the following tasks:

- **Image captioning:** Gemma3 [24], Qwen3-VL [25] generate descriptive captions from images, enabling semantic interpretation through language representations.
- **Image generation:** Our customized models based on StyleGAN3 [15], SPADE [20] and U-Net based discriminator [23] produces synthetic images to augment the dataset.
- **Image segmentation:** UNet [22], UNet++ [30], PAN [16], DeepLabV3+ [5], SegFormer [27] and FPN [17] generate semantic segmentation maps for scene understanding.
- **Composition description:** Gemma3 [24] and Qwen3 [25] convert scene composition into textual descriptions.

4. ARAS400k

The proposed framework for ARAS400k (A large-scale Remote sensing dataset Augmented with Synthetic data for segmentation and captioning), combines the strengths of vision and language domains. Specifically, semantic segmentation to process remote sensing images, providing structural representation of images with predetermined land cover classes. Composition of a scene from the outputs of the semantic segmentation is obtained to present a high-level summary. Remote sensing images, segmentation maps and compositions are used by language models to generate captions.

To demonstrate the effectiveness of our framework, we collected, cleaned and pre-processed remote sensing image and land cover data, as detailed in Section 3.1 and shown in Figure 2. Then we trained segmentation models from real image-segmentation pairs and generative models from real images. We utilized our generative models to generate synthetic samples, as shown in Figure 5, effectively quadrupling the volume of the existing dataset and segmentation models to obtain segmentation maps. Finally, we used pre-trained open-sourced foundation models for descriptive caption generation, using only images, composition statistics or both.

We used CLIPScore [11], the reference-free caption evaluation metric, to assess the alignment of the generated captions. We also calculated caption redundancy of the existing and our datasets, by finding unique strings in each dataset.

To improve downstream tasks, we introduced synthetic samples for segmentation model training to demonstrate the usefulness of the synthetic samples. This synthetic data augmentation effectively improved segmentation performance for all class-based and overall performance metrics. To prevent data-leakage, and enable reproducibility, we fixed train, validation and test sets and used only train set to generate synthetic samples.

Table 1. Comparison of remote sensing image caption datasets in terms of total number of images (volume), captions and redundancy.

Dataset	Volume	Captions	Redundancy	CLIPScore
NWPU	31,500	157,500	72.65%	30.25
RSICD	10,921	54,605	67.02%	29.11
UCMC	2,100	10,500	80.88%	30.18
ARAS400k	400,240	2,001,200	12.85%	29.66
-Real	100,240	501,200	8.01%	29.89
-Synth	300,000	1,500,000	13.06%	29.58

As shown in Table 1, the volume of our dataset, with 100,240 real and 300,000 synthetic images and a total of 2,001,200 captions, significantly exceeds existing remote sensing benchmarks such as NWPU [7] containing 31,500, RSICD [19] with 10,921, and UCMC [21] with 2,100. Furthermore, our dataset is much richer in caption variety. While traditional datasets like UCMC and NWPU show high redundancy rates of 80.88% and 72.65% respectively, ARAS400k achieves a significantly lower redundancy of 12.85%. This demonstrates that while previously introduced remote sensing caption datasets facilitate image-to-text training, they are relatively limited in scale and often contain repetitive captions introduced by human annotators.

Our captioning approach employs three distinct modalities using foundation models.

- Text: Converts segmentation-derived composition statistics into descriptions.
- Vision: Generates captions directly from image content.
- Hybrid: Integrates both visual content and composition statistics. Employs vision-language foundation models to generate captions and language models to refine percentage based composition statistics obtained from the segmentation maps.

Detailed comparisons in Table 2 show that hybrid methods achieve better performance and the lowest redundancy across both real and synthetic data subsets. While real-data subsets exhibit lower baseline redundancy, all configurations follow a consistent pattern: hybrid models yield the highest variety, whereas text-only result in the highest redundancy.

We also provide image segmentation models trained with the image-segmentation pairs. Finally, we trained generative models for synthetic data augmentation to increase the volume and variety of the existing real data. We used our segmentation models to obtain segmentation maps of our

Table 2. Captions generated by each model, input modality, unique sentences and the redundancy rates in our dataset.

Model	Method	Unique	Redundancy	CLIPScore
ARAS400k				
Qwen3-4B	Text	214,319	46.45 %	26.34
Gemma3-4B	Vision	396,878	0.84 %	31.06
Qwen3-VL-8B	Vision	359,997	10.05 %	31.79
Gemma3-4B	Hybrid	398,339	0.47 %	30.26
Qwen3-VL-8B	Hybrid	374,570	6.41 %	27.88
Real 100k				
Qwen3-4B	Text	71,899	28,27 %	26.48
Gemma3-4B	Vision	99,981	0.26 %	30.60
Qwen3-VL-8B	Vision	91,269	8.95 %	30.67
Gemma3-4B	Hybrid	100,105	0.13 %	29.90
Qwen3-VL-8B	Hybrid	97,814	2.42 %	27.87
Synthetic 300k				
Qwen3-4B	Text	159,138	46.95 %	26.29
Gemma3-4B	Vision	296,900	1.03 %	31.21
Qwen3-VL-8B	Vision	268,821	10.39 %	32.16
Gemma3-4B	Hybrid	298,556	2.70 %	30.38
Qwen3-VL-8B	Hybrid	280,731	6.42 %	27.88

synthetic data and obtained composition statistics, similar to real data as seen from the Figure 3.

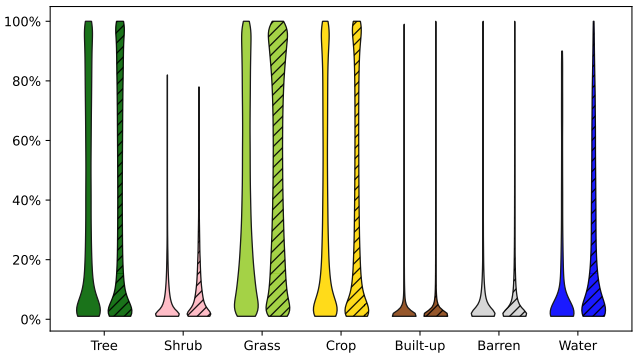

Figure 3. Real and synthetic (hatched) per-class composition statistics, indicating strong similarity between real and synthetic data.

We also compared t-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) of real and synthetic samples, using our trained segmentation model's feature extraction capability. The tightly grouped projection of real and synthetic samples inside similar clusters, as shown in the Figure 4, indicate our synthetic samples are visually similar to the real data.

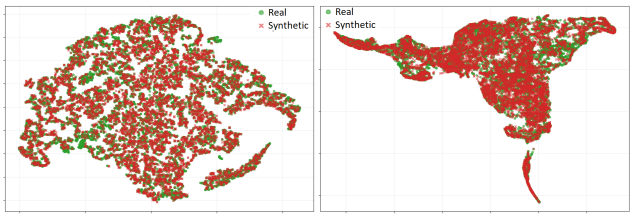

Figure 4. t-SNE (left) and UMAP (right) comparison of real and synthetic samples, showing similarity of real and synthetic data.

5. Experimental Evaluation

To ensure experimental rigor and prevent data leakage, the generative models were trained exclusively on a fixed training partition containing 80,182 real samples. Utilizing our implementation of the StyleGAN3 [15], U-Net based discriminator [23] and SPADE [20] architectures, we trained our GAN models until their training FID score reached a plateau, indicating that the model had converged. Our updates to the StyleGAN3 training code-base, utilizing newer introductions to the libraries, enabled 44% increase in the training and inference speed over the original official implementation. We present randomly selected real and synthetic samples in Figure 5, chosen for their visual similarity based on feature extraction from our pre-trained segmentation model. These comparisons show that the synthetic images are qualitatively similar to the real samples. FID scores for unconditional and conditional synthetic samples were measured as 16 and 72 respectively, indicating unconditional generation creates more realistic samples than the semantic segmentation mask conditioned generation. Furthermore, as shown in Table 1, existing datasets achieve CLIPScores between 29.11 and 30.25, meanwhile ARAS400k scores 29.89 and 29.58 CLIPScore for real and synthetic subsets respectively, indicating highly-competitive performance with human-annotated benchmarks.

Table 3 presents the macro average F1 score, Intersection over Union (IoU), precision, recall and accuracy across all experimental configurations. The results of the models trained only with synthetic data are comparable yet slightly inferior to real data, indicating synthetic-only variants of our dataset is still capable to train solid models, with larger (300k) variant achieving better results than the smaller (100k). The augmented data configurations, consisting of both real and synthetic samples, achieve better results than the real data baseline, clearly indicating the advantage of synthetic data augmentation. Utilizing real data augmented with unconditional synthetic data samples (Real + Uncond. 300k) performs better than augmentation with conditional synthetic data (Real + Cond. 80k) or with both (Real + Uncond. 300k + Cond. 80k).

Table 4 presents a comprehensive per-class evaluation, reporting F1 scores for each land cover class. The results

Table 3. Overall test performance comparison of segmentation architectures trained on real and synthetic data (values in %).

	Real Data					**Uncond. Synthetic 100k**					**Uncond. Synthetic 300k**				
Model	F1	IoU	Prec.	Recall	Acc.	F1	IoU	Prec.	Recall	Acc.	F1	IoU	Prec.	Recall	Acc.
Unet	76.37	64.49	77.05	75.86	85.17	73.06	60.79	**78.17**	69.71	83.24	74.63	62.31	77.60	72.23	83.85
Unet++	76.38	64.36	76.50	76.39	84.85	**73.57**	**61.27**	77.84	**70.57**	83.35	74.94	62.73	77.81	72.56	84.28
PAN	76.20	64.10	75.29	**77.60**	84.62	73.07	60.71	77.53	70.04	83.19	74.65	62.42	77.70	72.22	84.14
DeepLab	76.02	63.91	75.59	76.77	84.37	72.43	59.91	77.08	69.23	82.51	74.63	62.39	77.69	72.15	84.09
Segformer	**77.09**	**65.26**	**78.10**	76.33	**85.66**	73.48	61.23	77.77	70.40	**83.63**	**75.10**	**62.90**	77.65	**72.95**	**84.33**
FPN	76.85	64.92	77.07	76.76	85.17	73.36	61.01	77.30	70.43	83.36	74.81	62.59	**77.99**	72.26	84.26
Mean	76.49	64.51	76.60	**76.62**	84.97	73.16	60.82	77.62	70.06	83.21	74.79	62.56	77.74	72.40	84.16

	Real + Uncond. 300k + Cond. 80k					**Real + Cond. 80k**					**Real + Uncond. 300k**				
Model	F1	IoU	Prec.	Recall	Acc.	F1	IoU	Prec.	Recall	Acc.	F1	IoU	Prec.	Recall	Acc.
Unet	76.36	64.52	79.41	74.00	85.43	78.00	66.39	79.11	**77.04**	86.20	77.40	65.58	79.37	75.64	85.79
Unet++	**77.70**	**66.04**	**79.70**	**75.96**	86.08	**77.98**	**66.40**	**79.52**	76.70	86.19	**78.23**	**66.67**	80.05	**76.65**	**86.44**
PAN	77.29	65.52	79.21	75.64	85.86	76.00	63.90	77.00	75.17	84.50	77.21	65.48	79.94	75.02	86.04
DeepLab	77.18	65.42	79.48	75.27	85.91	77.59	65.87	78.33	76.91	85.83	77.62	65.94	80.00	75.65	86.20
Segformer	77.43	65.72	79.58	75.61	**86.10**	77.91	66.27	79.09	76.83	**86.20**	77.80	66.14	79.98	75.98	86.30
FPN	77.45	65.74	79.51	75.68	86.08	77.88	66.30	79.57	76.43	86.29	77.71	66.04	**80.16**	75.66	86.28
Mean	77.24	65.49	79.48	75.36	85.91	77.56	65.86	78.77	76.51	85.87	**77.66**	**65.98**	**79.92**	75.77	**86.18**

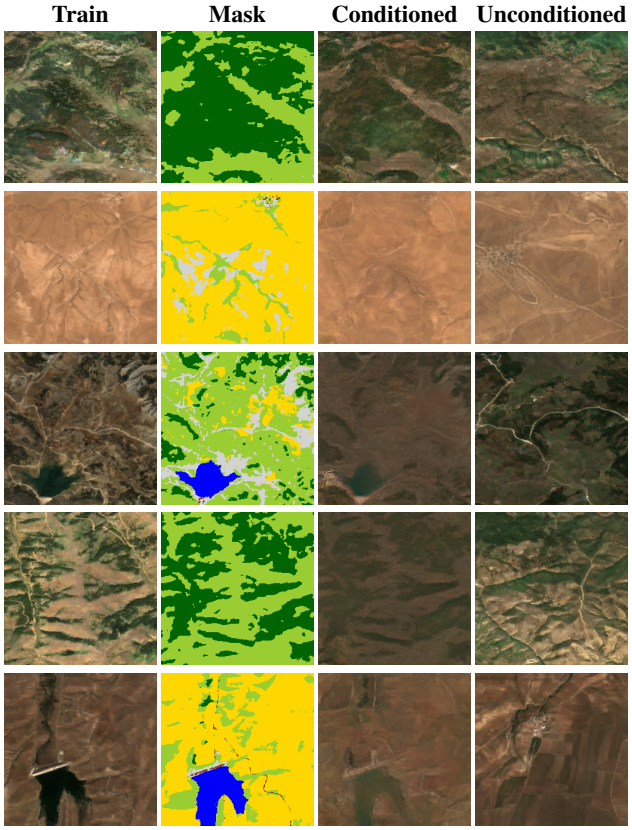

Figure 5. Qualitative comparison of training images, masks, conditioned and unconditioned synthetic samples.

reveal substantial performance variation across categories: Water class achieves the highest scores, while shrubs pose a significant segmentation challenge, possibly due to scarcity of the shrub class and similarity with neighboring classes. Tree cover, croplands and grasslands were also successfully distinguished by the segmentation models, followed by built-up and barrens. Synthetic data provides comparable but slightly inferior results compared to the real data meanwhile the combination of real and synthetic data, shown as augmented data, achieves better performance across all classes consistently. The highest performance increase occurs in the least represented and lowest performing classes, indicating that synthetic data augmentation effectively alleviates class-imbalance problem.

6. Discussion

We trained semantic segmentation models with real images, using different architectures, creating a baseline performance benchmark for this downstream task. To evaluate the contribution of conditional and unconditional synthetic data augmentation approaches, we used this baseline and re-trained same architectures with synthetic-augmented data. Details of this ablation study, shown in the Table 3, highlight that using synthetic only data is an alternative to real data at the cost of a slight performance decrease. Using a comparable number of synthetic samples as the original real data, decreases the overall F1 score by 3.5 points. Meanwhile utilizing 3 times larger synthetic samples achieves a better result, trailing the real baseline by 2.0 F1 score. These results show that using synthetic samples instead of real data is feasible at a slight

Table 4. Per-class F1-score comparison of segmentation models (values in %).

	Real Data							**Uncond. Synthetic 100k**						
Model	Tree	Shrub	Grass	Crop	Build	Barren	Water	Tree	Shrub	Grass	Crop	Build	Barren	Water
Unet	90.92	45.86	84.25	86.82	73.86	58.98	93.88	89.80	35.65	82.70	83.93	71.64	55.84	91.88
Unet++	90.88	45.50	83.95	86.30	**74.11**	60.97	92.93	**90.04**	**37.64**	82.82	83.90	**71.85**	**56.66**	92.05
PAN	90.31	46.88	84.05	86.38	71.50	60.58	93.69	89.73	36.07	82.74	83.90	70.94	56.57	91.54
DeepLab	90.42	46.22	83.56	86.24	72.21	59.92	93.58	89.61	36.40	81.97	82.66	70.13	54.76	91.50
Segformer	**91.08**	**48.80**	**85.07**	**87.16**	73.05	60.25	**94.20**	90.04	37.40	**83.08**	**84.51**	71.13	56.14	**92.06**
FPN	90.40	47.67	84.48	87.12	73.02	**61.21**	94.05	89.87	38.01	82.85	84.11	71.01	55.74	91.90
Mean	90.67	46.82	84.23	86.67	72.96	60.32	93.72	89.85	36.86	82.69	83.83	71.12	55.95	91.82

	Uncond. Synthetic 300k							**Real + Uncond. 300k + Cond. 80k**						
Model	Tree	Shrub	Grass	Crop	Build	Barren	Water	Tree	Shrub	Grass	Crop	Build	Barren	Water
Unet	90.31	42.23	83.33	84.53	**72.16**	57.61	92.22	91.02	45.44	84.64	86.75	74.03	58.87	93.75
Unet++	90.40	42.73	83.73	**85.38**	72.05	57.85	92.44	91.47	**48.17**	85.21	87.56	**74.74**	**62.54**	**94.19**
PAN	90.34	41.73	83.61	85.18	71.83	57.63	92.25	91.29	48.06	85.04	87.36	73.50	61.74	94.03
DeepLab	90.35	41.71	83.56	85.05	71.90	57.58	92.27	91.39	47.82	85.13	87.30	73.69	60.92	94.01
Segformer	90.43	**43.35**	**83.85**	85.37	72.06	**58.01**	**92.63**	91.47	48.07	**85.29**	**87.58**	73.59	61.87	94.12
FPN	**90.44**	42.39	83.76	85.19	71.81	57.71	92.39	**91.50**	**48.17**	85.19	**87.58**	73.64	61.94	94.11
Mean	90.38	42.36	83.64	85.12	71.97	57.73	92.37	91.36	47.96	85.08	87.36	73.87	61.31	94.04

	Real + Cond. 80k							**Real + Uncond. 300k**						
Model	Tree	Shrub	Grass	Crop	Build	Barren	Water	Tree	Shrub	Grass	Crop	Build	Barren	Water
Unet	91.58	48.94	85.41	87.58	74.92	63.14	94.42	91.42	48.59	84.89	87.23	73.66	62.29	93.69
Unet++	91.59	48.35	85.46	87.54	**75.36**	**63.27**	94.32	**91.77**	**49.75**	**85.61**	**87.83**	**75.18**	**63.18**	**94.31**
PAN	90.25	45.69	83.26	86.35	72.31	60.35	93.78	91.38	47.61	85.24	87.53	73.20	61.53	94.00
DeepLab	91.36	48.81	85.05	87.24	73.95	62.26	94.43	91.53	48.88	85.37	87.69	73.79	61.93	94.15
Segformer	91.58	**49.20**	85.47	**87.65**	74.11	62.97	94.36	91.57	48.99	85.48	87.80	73.75	62.85	94.16
FPN	**91.72**	48.67	**85.61**	87.54	74.31	62.62	**94.65**	91.55	48.94	85.51	87.74	73.49	62.54	94.19
Mean	91.35	**48.94**	85.04	87.32	**74.16**	**62.44**	**94.33**	**91.54**	48.79	**85.35**	**87.64**	73.85	62.05	94.08

performance cost.

Injection of synthetic samples to real data improves the overall segmentation performance over the baseline. We tested three augmentation strategies: augmentation with segmentation map conditioned synthetic data, with unconditioned synthetic data, and a combination of both. Augmentation with unconditioned synthetic data slightly outperforms other strategies. The most important gain in using synthetic augmentation is more evident with under-represented and under-performing classes, as evident in the Table 4. These results indicate that synthetic data augmentation improves downstream performance for the rare cases.

The results of these evaluations, summarized in the Table 1, indicate that our approach is much more diverse than existing datasets, while having between $12-190\times$ more samples. Although CLIPScore for existing datasets are slightly higher than our dataset, some of our models performed better. As shown in the Table 2, vision models tend to provide more realistic descriptions in terms of CLIPScore. One distinct feature of our synthetic dataset over existing human-annotated datasets is the presence of land cover percentages. Our caption generation models, except for vision-only methods, had access to the composition statistics of the scene. We believe including this percentage information for generated captions is a valuable contribution, albeit the negative affects for the CLIPScore evaluation metric.

On a single NVIDIA H100 GPU, total processing and ablation study for this dataset spanned over 3,000 hours. A single end-to-end execution requires 80 hours for generative model training and synthetic sample generation, 20 hours for segmentation model training and inference, 300 hours for caption generation of 5 captions per each of 400,000 samples, totaling approximately 400 hours for whole dataset. Our framework processes 1,000 images per hour.

7. Conclusion

We introduced a vision-language framework for remote sensing synthetic data augmentation that integrates generative modeling, semantic segmentation, and automated image captioning. Using this framework, we produced ARAS400k, a

large-scale dataset featuring over 100k real and 300k synthetic image-segmentation pairs. By utilizing foundation models to generate captions from both visual content and segmentation-derived composition statistics, our approach provides a scalable and interpretable solution for enhancing multi-modal remote sensing tasks.

The experimental results for semantic segmentation establish a robust baseline with real data, while models trained exclusively on synthetic data achieved highly competitive performance. Most notably, models trained on augmented data (real and synthetic) achieved superior results, demonstrating the advantage of synthetic augmentation for scalable data expansion. Meanwhile reference-free CLIPScore evaluation confirm that our automated captions are comparable to human-annotated datasets while having significantly higher volume and lower redundancy. The consistent performance across real and synthetic subsets highlights the potential for using synthetic data as a viable alternative to sensitive or labor-intensive real-world datasets.

Possible future work includes improving synthetic image quality through remote sensing super-resolution approaches, developing specialized evaluation metrics for synthetic samples and exploring structure preserving image generation from segmentation maps. We have introduced this work as an example in remote sensing, due to the availability of large scale data acquisition through public data sources and we believe this work can be replicated for different domains, including but not limited to common scene understanding, autonomous driving and medical domains.

Acknowledgment

Ümit Mert Çağlar is a beneficiary of the TUBITAK BIDEB 2211-A Domestic Graduate Scholarship Program. The experiments reported in this work were fully performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA).

References

[1] Ahmad Al-Qerem, Ali Mohd Ali, Hani Attar, Shadi Nashwan, Lianyong Qi, Mohammad Kazem Moghimi, and Ahmed Solyman. Synthetic generation of multidimensional data to improve classification model validity. *ACM Journal of Data and Information Quality*, 15(3):1–20, 2023. 2

[2] Ajay Bandi, Pydi Venkata Satya Ramesh Adapa, and Yudu Eswar Vinay Pratap Kumar Kuchi. The power of generative ai: A review of requirements, models, input–output formats, evaluation metrics, and challenges. *Future Internet*, 15(8): 260, 2023. 1

[3] Farhat Lamia Barsha and William Eberle. An in-depth review and analysis of mode collapse in generative adversarial networks. *Machine Learning*, 114(6):141, 2025. 2

[4] Eyal Betzalel, Coby Penso, and Ethan Fetaya. Evaluation metrics for generative models: An empirical study. *Mach. Learn. Knowl. Extr.*, 6(3), 2024. 1

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 3

[6] Yunhao Chen, Zihui Yan, and Yunjie Zhu. A comprehensive survey for generative data augmentation. *Neurocomputing*, 600:128167, 2024. 1

[7] Qimin Cheng, Haiyan Huang, Yuan Xu, Yuzhuo Zhou, Huanying Li, and Zhongyuan Wang. Nwpu-captions dataset and mlca-net for remote sensing image captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2022. 3

[8] Xu Guo and Yiqiang Chen. Generative ai for synthetic data generation: Methods, challenges and the future. *arXiv:2403.04190*, 2024. 1

[9] Hasan Abed Al Kader Hammoud, Hani Itani, Fabio Pizzati, Philip Torr, Adel Bibi, and Bernard Ghanem. Synthclip: Are we ready for a fully synthetic clip training?, 2024. 2

[10] Ruozhen He, Ziyan Yang, Paola Cascante-Bonilla, Alexander C. Berg, and Vicente Ordonez. Learning from synthetic data for visual grounding, 2024. 2

[11] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Conference on Empirical Methods in Natural Language Processing*, 2021. 2, 3

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 2020. 2

[13] Nick Huang, Aaron Gokaslan, Volodymyr Kuleshov, and James Tompkin. The GAN is dead; long live the GAN! a modern GAN baseline. *Advances in Neural Information Processing Systems*, 37:44177–44215, 2024. 2

[14] Minguk Kang, Richard Zhang, Connelly Barnes, Sylvain Paris, Suha Kwak, Jaesik Park, Eli Shechtman, Jun-Yan Zhu, and Taesung Park. Distilling diffusion models into conditional GANs. In *ECCV*, 2024. 2

[15] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *International Conference on Neural Information Processing Systems*, 2021. 2, 3, 4

[16] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. *arXiv:1805.10180*, 2018. 3

[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] Xinfu Liu, Benze Wu, and Yirui Wu. A remote sensing image classification method based on detail attention sampling and teacher-student network. *ACM Journal of Data and Information Quality*, 17(3):1–19, 2025. 2

[19] Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195, 2017. 3

[20] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2332–2341, 2019. 3, 4

[21] Bo Qu, Xuelong Li, Dacheng Tao, and Xiaoqiang Lu. Deep semantic understanding of high resolution remote sensing image. In *2016 International Conference on Computer, Information and Telecommunication Systems (CITS)*, pages 1–5, 2016. 3

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[23] Edgar Schonfeld, Bernt Schiele, and Anna Khoreva. A U-net based discriminator for generative adversarial networks. In *CVPR*, 2020. 3, 4

[24] Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. 3

[25] Qwen Team. Qwen3 technical report, 2025. 3

[26] Shaobo Wang, Yantai Yang, Qilong Wang, Kaixin Li, Linfeng Zhang, and Junchi Yan. Not all samples should be utilized equally: Towards understanding and improving dataset distillation. *arXiv preprint arXiv:2408.12483*, 2024. 2

[27] 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. 3

[28] Shamim Yazdani, Akansha Singh, Nripsuta Saxena, Zichong Wang, Avash Palikhe, Deng Pan, Umapada Pal, Jie Yang, and Wenbin Zhang. Generative ai in depth: A survey of recent advances, model variants, and real-world applications. *Journal of Big Data*, 12(1):230, 2025. 2

[29] Daniele Zanaga, Ruben Van De Kerchove, Dirk Daems, Wanda De Keersmaecker, Carsten Brockmann, Grit Kirches, Jan Wevers, Oliver Cartus, Maurizio Santoro, Steffen Fritz, et al. Esa worldcover 10 m 2021 v200, 2022. 2

[30] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *International workshop on deep learning in medical image analysis*, 2018. 3