IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE GEOSCIENCE AND REMOTE SENSING SOCIETY SECTION

Received 25 October 2025, accepted 10 November 2025, date of publication 12 November 2025,
date of current version 18 November 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3632152

RESEARCH ARTICLE

Multispectral Image Caption Unification Using Diffusion and Cycle GAN Models

KURSAT KOMURCU AND LINAS PETKEVICIUS, (Member, IEEE)
Institute of Computer Science, Faculty of Mathematics and Informatics, Vilnius University, 08303 Vilnius, Lithuania

Corresponding author: Kursat Komurcu (kursat.komurcu@mif.vu.lt)

This work was supported by European Union through the Research Council of Lithuania [Lietuvos Mokslo Taryba (LMT)] under Project S-MIP-23-45.

ABSTRACT A major limitation is the scarcity of geospatial datasets that simultaneously provide multispectral imagery and descriptive captions. In particular, datasets containing aligned RGB, multispectral, and caption information remain highly limited. Therefore, we propose a full-circle pipeline to unify triplets of RGB images, image captions, and Sentinel-2–like multispectral data. To accomplish this, we combine a fine-tuned Stable Diffusion model with a Cycle GAN trained on generated images and the EuroSAT dataset. First, we use Qwen2-VL-2B as a zero shot method to generate captions for 675,993 images from the SkyScript dataset. We then fine-tune the Stable Diffusion 2-1 Base model on these image–caption pairs and generate randomly selected 123,081 RGB images conditioned on the Qwen2-VL-2B captions. Finally, we train a Cycle GAN on roughly 27,000 paired RGB and multispectral images and use it to translate synthetic RGB images into multispectral counterparts. In this way, textual prompts produce synthetic satellite imagery that can be converted to multispectral Sentinel-2 data. The pipeline enables unifying datasets that contain only captions or only RGB images by producing complete triplets (caption, RGB, multispectral). Quantitative evaluations support the credibility of the approach: generated captions achieve a SkyCLIP Score of 0.7312, the fine-tuned Stable Diffusion model achieves a CMMD of 0.245, and the Cycle GAN multispectral outputs reach a SAM of 10.16° in our synthetic dataset versus 13.94° on EuroSAT. The code, models and the dataset links are available at GitHub and Hugging Face.

INDEX TERMS Cycle GAN, data synthesis, image-to-image translation, multispectral imagery, remote sensing, sentinel-2, stable diffusion, text-to-image.

I. INTRODUCTION

Computer vision applications on RGB imagery have led to significant breakthroughs in the geospatial domain in recent years [1], [2], [3], [4]. However, the availability of labeled multispectral datasets (e.g. Sentinel-2 [5]) remains limited. Although datasets exist for RGB object detection, semantic segmentation, and image captioning, the largest open-source multispectral sources, such as Sentinel-2 [5], rarely include captions.

Generative models, particularly diffusion-based approaches and generative adversarial networks (GANs), have demonstrated remarkable capabilities in image synthesis and transformation [6]. However, generating high-quality multispectral satellite imagery from textual descriptions remains challenging due to the complex spectral characteristics inherent in remote sensing data [7].

Existing research has explored the use of Cycle GAN for image-to-image translation in remote sensing and geospatial applications [8]. Furthermore, diffusion models have recently gained prominence in hyperspectral image synthesis, demonstrating the potential to generate high-fidelity data with improved spatial and spectral consistency [9]. However, integrating diffusion models with a Cycle GAN for caption-conditioned multispectral image generation remains underexplored.

In this study, we address the problem of unifying existing datasets of captions, RGB images, and Sentinel-2 like multispectral data (later we will call it a triplet). We propose a novel methodology that aims to integrate

The associate editor coordinating the review of this manuscript and approving it for publication was Donato Amitrano.

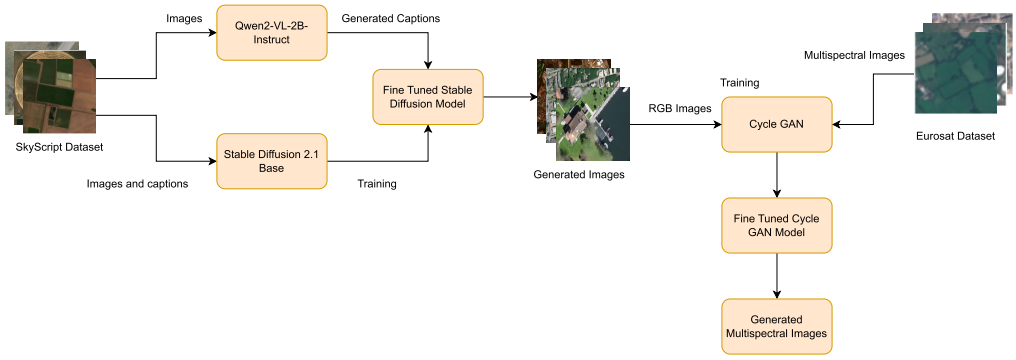

FIGURE 1. Workflow of the proposed methodology. Qwen2-VL-2B-Instruct generates captions in a zero-shot manner, which are then used by a Stable Diffusion model (fine-tuned on original RGB–caption pairs) to synthesize new RGB images. These generated RGB images are subsequently translated into Sentinel-2-like multispectral images using a Cycle GAN, which was pre-trained on paired original RGB and Sentinel-2 data. The pipeline enables the construction of unified triplets of caption, RGB, and multispectral imagery.

caption-based synthetic image generation with multispectral image translation techniques. By combining Stable Diffusion models with Cycle GAN based models, we create a pipeline that enables the generation of realistic synthetic missing data of the triplet. This approach allows us to generate the missing data of the triplet and propose a unified process to combine the datasets that could not be combined to date.

The key motivation of our work is to address the unifying different data types in remote sensing. Although diffusion models are good at generating semantically coherent images from textual descriptions, they cannot capture fine-grained spectral distributions. However, we thought that Cycle GAN could be effective in translating RGB to multispectral domains, thereby encoding spectral relationships. To the best of our knowledge, only limited attempts have been made to explicitly explore the RGB-to-Sentinel-2 translation. By integrating these complementary strengths, our method directly tackles the challenge of spectral complexity, enabling the generation of semantically meaningful and spectrally consistent multispectral imagery.

This research contributes to the field by:

- Proposal for a pipeline to unify caption, RGB, and multispectral Sentinel-2 image datasets.
- Delivering the Stable Diffusion 2-1 Base model to improve text-to-image generation, as well as creating Cycle GAN [10] to convert RGB images to multispectral Sentinel-2 images, and vice versa.
- Creating a new artifical multispectral satellite image-caption dataset.

By unifying these techniques, our study aims to improve the applicability of remote sensing imagery generated. This approach allows to expand accessible data and successfully improve remote sensing applications such as land cover classification, disaster monitoring, and environmental change detection [11].

II. RELATED WORK

Existing remote sensing (RS) caption datasets are predominantly RGB-only (e.g., RSICD and NWPU-Captions), while widely used multispectral (MSI) resources (e.g., EuroSAT / Sentinel-2) typically lack human captions [5], [12], [13], [14]. Recent efforts have begun to pair Sentinel-2 imagery with LLM-generated text at global scale (e.g., ChatEarthNet) or to scale up RS vision–language corpora (e.g., SkyScript), yet these corpora still provide image–text pairs rather than co-registered caption–RGB–MSI triplets and do not offer a mechanism to translate across modalities [15], [16].

Generative models have contributed significantly to computer vision [17], [18] and remote sensing [19], [20], particularly in image synthesis and transformation. Among these, Cycle GAN has shown effectiveness in image-to-image translation tasks [10]. It has been widely used in geospatial analysis and remote sensing, where it has been used for domain adaptation and multispectral image synthesis [8]. Recent studies have improved Cycle GAN's image translation accuracy on Sentinel-2 remote sensing data [21].

Similarly, style transfer between cartographic maps and satellite images has been attempted to translate a city street map into a pseudo-satellite image of that city, and vice versa [22]. The result is a synthetic image that looks like a Sentinel or Google Earth view of a city given only the map. Extensions of Cycle GAN, such as AttentionGAN [23], have also been tried to address the focus on important regions or to preserve edges. Although diffusion models and Cycle GAN have been widely explored individually, their combined potential for multispectral image synthesis remains underexplored. A novel study [24] has applied GAN-based approaches to multispectral images using two SAR images and a RGB image.

Recently, large-vision language models have been explored for captioning. An approach is to use a pre-trained vision encoder such as a CLIP [25] visual backbone or a ViT [26] trained on ImageNet and connect it to a pre-trained

TABLE 1. Comparison of pipeline with other studies.

Method / Resource	Cap→RGB	RGB→MSI	Triplet-ready
Stable Diffusion [29]	✓	–	–
Pix2Pix [30]	–	✓	–
CycleGAN [8], [10]	–	✓	–
Hyperspectral diffusion [9]	–	✓†	–
Ours	✓	✓	✓

TABLE 2. Comparison of our dataset with other datasets.

Dataset	Caption	RGB	Multispectral
RSICD [12]	✓	✓	–
RS5M [32]	✓	✓	–
Eurosat [14]	–	✓	✓
SkyScript [16]	✓	✓	–
ChatEarthNet [15]	✓	✓	✓
Ours	✓	✓	✓

language model. Recently, the Geochat model [27] was proposed as a vision language model for remote sensing.

In addition to the studies mentioned above, our experimental setup also involves several representative baseline models. For caption generation, we employ Qwen [28] as a recent vision–language model capable of producing descriptive captions from images. For text-to-image generation, The Stable Diffusion 2-1 Base model [29] serves as a state-of-the-art diffusion-based model, which we fine-tune on original captions and RGB images to synthesize realistic RGB imagery. For image-to-image translation, Cycle GAN [10] is used as the baseline to translate RGB imagery into multispectral Sentinel-2 data, which we trained using original RGB and multispectral images. This work offers a preliminary attempt at RGB to Sentinel-2 translation, a direction for which we found little prior research. Rather than claiming to fully solve the challenge, our aim is to demonstrate the feasibility of such a translation within a unified framework. Table 1 positions our pipeline against closest baselines (caption→RGB, RGB→MSI, triplet readiness), while Table 2 contrasts dataset modality coverage and shows that our release provides aligned caption–RGB–MSI triplets.

III. DATASET

In this study, we used multiple datasets to facilitate the training and evaluation of our proposed methodology. We used two datasets: the SkyScript dataset [16], the Eurosat dataset [14], [32]. We also created our synthetic dataset. Each dataset serves a distinct purpose in training the Stable Diffusion and Cycle GAN models.

The SkyScript dataset [16] comprises 5.2M satellite images accompanied by captions. We use this dataset to generate captions using the Qwen2-VL-2B-Instruct model [28] and compare them with the original captions using 675,993 images. The reason why we chose this model is the ''Performance Comparison of Qwen2-VL Models and State-of-the-art'' Table in the paper [28]. These original image-caption pairs serve as the primary training data for fine-tuning our Stable Diffusion 2-1 base model. Using this dataset, we aim to enhance the ability of our model to generate realistic satellite imagery from textual descriptions while maintaining semantic consistency, and we used our generated captions to generate images after the Stable Diffusion 2-1 base model training.

To train our Cycle GAN model, we used 27,000 RGB and multispectral images from the Eurosat dataset [14], [32]. This dataset contains RGB images and their corresponding 13-band Sentinel-2 multispectral representations, which are essential for training the Cycle GAN model to perform accurate image-to-image translation. The Eurosat dataset is widely used in remote sensing applications and provides multispectral data for various geospatial tasks [14], [32]. Using this dataset, we ensure that our model learns to transform RGB images into realistic multispectral representations, and we use our generated RGB images to create multispectral images after the Cycle GAN training.

In our experiments, 123,081 images from the Skyscript dataset, which we had previously randomly selected and captioned, were used to generate multispectral images. The novel synthetic dataset serves multiple purposes: first, it allows us to extend the dataset across all three modalities; secondly, we present an initial benchmark for the translation of RGB to Sentinel-2, offering a resource that may support future studies in the field of remote sensing.

Moreover, we must note that neither EuroSAT nor SkyScript provide per-image sun/view geometry; patches are cropped from heterogeneous Sentinel-2 scenes or web-scale RGB imagery, so illumination and viewing conditions vary across samples and are not controlled in our training/test splits.

IV. METHODOLOGY

The proposed methodology aims to unify caption, RGB, and multispectral data with the ability to translate all three types of data. Our workflow consists of two main stages: text-to-image generation by fine-tuning the Stable Diffusion model and image-to-multispectral conversion using a Cycle GAN model; see Fig. 1. This proposed pipeline allows the creation of realistic synthetic satellite imagery that can be transformed into multispectral Sentinel-2 representations. The experiments were implemented using the Google Colab Pro+ account and the A100 GPU was used during this study.

The Qwen2-VL-2B-Instruct model [28] was used in our experiments as a zero shot due to its performance in large-scale datasets and its low size. These generated captions were compared with the original captions provided with the dataset. Comparison was performed using both automated text similarity metrics like widely used evaluation metrics, including BLEU [33], METEOR [34], ROUGE-L [35], CIDEr-D [36], BERT-F1 [37], and CLIP Scores [38] which we calculated using CLIP [25], RemoteCLIP [39], and SkyCLIP [16] models used to evaluate the quality and semantic accuracy of the generated captions.

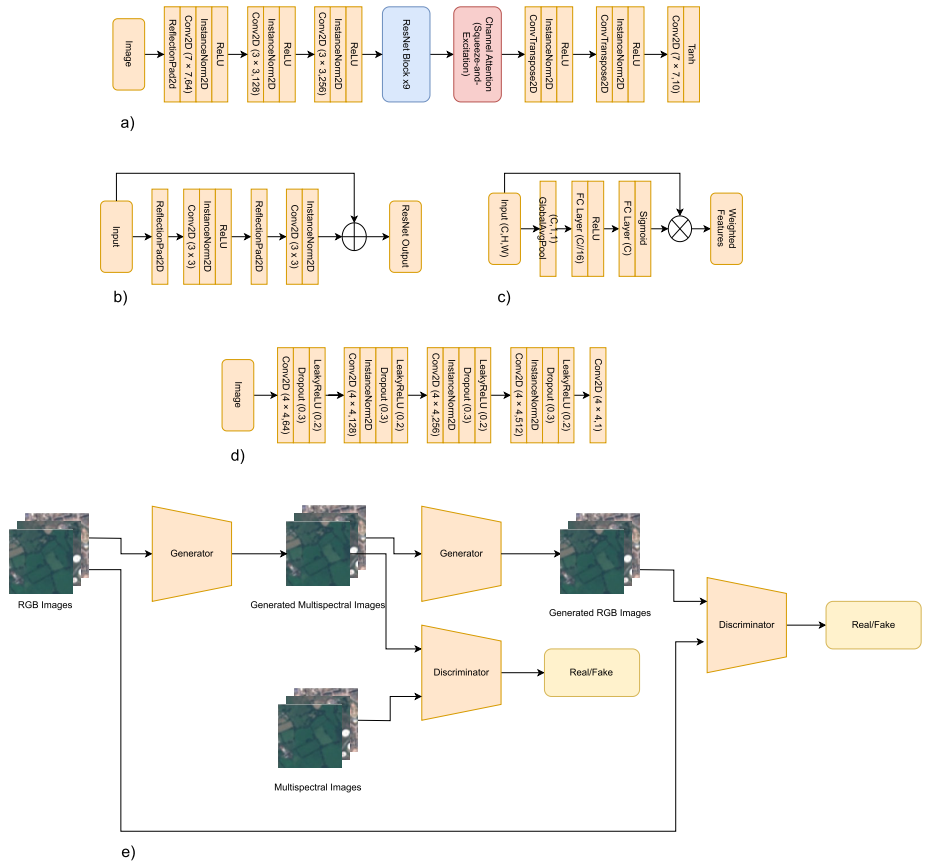

FIGURE 2. a) Generator Architecture, b) Discriminator Architecture, c) Channel Attention (Squeeze & Excitation) Block, d) ResNet Block, e) Cycle GAN diagram.

The Stable Diffusion 2-1 Base model [29] was used in our experiments and was chosen for its high image generation quality for text-to-image synthesis [40], [41]. The pretrained weights of the model made it an ideal candidate for fine-tuning on the domain-specific SkyScript dataset. The fine-tuning process involved using image-caption pairs derived from the SkyScript dataset. Batch size of 4 and a learning rate of 1×10^{-5}, 1 epoch and MSE (Mean Squared Error) loss, were selected for the fine-tuning process. After fine-tuning, the model performance was assessed by generating 123,081 synthetic images.

We decompose the pipeline into text→RGB synthesis and RGB→MSI translation. For text→RGB synthesis, diffusion offers stable text-conditioned diversity under domain fine-tuning [29]. For RGB→MSI translation, although our training uses paired RGB–Sentinel-2 tiles (EuroSAT), we adopt a CycleGAN backbone with Spectral Angel Mapper (SAM) and histogram losses for two reasons: (i) cycle consistency provides a structure-preserving inductive bias and enables bidirectional RGB↔MSI use within our triplet pipeline [10]; (ii) it is empirically robust to small resolution mismatches that often remain even in nominally paired RS products [42].

When the Cycle GAN model is compared with alternative models such as Pix2Pix [30] or StarGAN [43], its advantage lies in its ability to handle transformations without the need for strictly paired training data, which is critical for converting RGB images to multispectral formats [44], [45], [46]. The Cycle GAN model was trained using approximately 27,000 RGB and multispectral images from the Eurosat dataset using 80% images as train and 20% as validation. After the training, we evaluated our model on both Eurosat dataset and our dataset. In this setup, RGB images served as inputs and Eurosat 13-band Sentinel-2 multispectral images acted as target outputs. The training process converts input RGB images (3 channels) into output multispectral images consisting of 13 spectral bands. This transformation is designed to map the RGB domain into the multispectral domain, preserving the critical spectral characteristics necessary for remote sensing applications.

The Cycle GAN model used in this study consists of a generator and a discriminator network, designed to translate RGB images into multispectral representations while preserving critical spectral information (Figure 2) and selected a batch size of 512, 100 epochs, and a learning rate of 2×10^{-4}. RGB inputs are converted to $[0, 1]$ and normalized to $[-1, 1]$ using per-channel mean 0.5 and std 0.5, Sentinel-2 13-band inputs receive the same uniform $[-1, 1]$ normalization across all bands, no resizing or data

augmentation is applied, and the pipeline assumes 64 × 64 inputs. The generator network G 4b takes input RGB images (3 channels, domain A) and maps them to the target multispectral space (13 channels, domain B) through a deep convolutional architecture enhanced with residual learning and channel attention mechanisms. Generator F has the opposite task.

Traditional Cycle GAN models typically enforce cycle consistency using pixel-based losses such as L1 or L2 norms. However, in the context of converting RGB images to multispectral representations, preserving the inherent spectral relationships between bands is crucial. To address this, we integrated the Spectral Angle Mapper (SAM) [47] and histogram loss [48] into our model. SAM loss calculates the angular differences of the spectral vector and provides a consistent measure of spectral similarity with minimal illumination intensity fluctuation sensitivity. We also utilize a histogram loss to align the global distribution of ground-truth spectral signatures and those generated to keep spectral features equal across the whole image. The approach ensures that resultant multispectral images possess the minimum spectral features to ensure accurate remote sensing analysis. The equation for SAM and Histogram Loss (1), (1a), (1b), (4), (4a), (4b), (4c), (5), (6), (7):

$$\mathcal{L}_{\mathrm{SAM}}(x,y) = \arccos\left(\frac{\langle x,y\rangle}{\|x\|_2\,\|y\|_2}\right), \tag{1a}$$

$$\mathcal{L}_{\mathrm{hist}}(S^+,S^-) = \sum_{b=1}^{B} h_b^- \sum_{a=1}^{b} h_a^+. \tag{1b}$$

where

$$h_b^+ = \frac{1}{|S^+|}\sum_{s^+\in S^+} \mathbf{1}\{\, s^+ \in \mathrm{bin}_b \,\}, \tag{2}$$

$$h_b^- = \frac{1}{|S^-|}\sum_{s^-\in S^-} \mathbf{1}\{\, s^- \in \mathrm{bin}_b \,\}. \tag{3}$$

$$\begin{aligned}\mathcal{L}_{\mathrm{GAN}} &= \mathbb{E}_{a\sim p_A}\big[(D_B(G(a))-1)^2\big] \\ &\quad + \mathbb{E}_{b\sim p_B}\big[(D_A(F(b))-1)^2\big]\end{aligned} \tag{4a}$$

$$\begin{aligned}\mathcal{L}_{\mathrm{cyc}} &= \mathbb{E}_{a\sim p_A}\big[\mathcal{L}_{\mathrm{SAM}}(F(G(a)),a)\big] \\ &\quad + \mathbb{E}_{b\sim p_B}\big[\mathcal{L}_{\mathrm{SAM}}(G(F(b)),b)\big]\end{aligned} \tag{4b}$$

$$\begin{aligned}\mathcal{L}_{\mathrm{hist,tot}} &= \mathbb{E}_{a\sim p_A}\big[\mathcal{L}_{\mathrm{hist}}(G(a),a)\big] \\ &\quad + \mathbb{E}_{b\sim p_B}\big[\mathcal{L}_{\mathrm{hist}}(F(b),b)\big]\end{aligned} \tag{4c}$$

$$\mathcal{L}_G = \mathcal{L}_{\mathrm{GAN}} + \lambda_{\mathrm{cycle}}\,\mathcal{L}_{\mathrm{cyc}} + \lambda_{\mathrm{hist}}\,\mathcal{L}_{\mathrm{hist,tot}}. \tag{5}$$

$$\begin{aligned}\mathcal{L}_{D_A} &= \frac{1}{2}\,\mathbb{E}_{a\sim p_A}\big[(D_A(a)-1)^2\big] \\ &\quad + \frac{1}{2}\,\mathbb{E}_{b\sim p_B}\big[D_A(F(b))^2\big].\end{aligned} \tag{6}$$

$$\begin{aligned}\mathcal{L}_{D_B} &= \frac{1}{2}\,\mathbb{E}_{b\sim p_B}\big[(D_B(b)-1)^2\big] \\ &\quad + \frac{1}{2}\,\mathbb{E}_{a\sim p_A}\big[D_B(G(a))^2\big].\end{aligned} \tag{7}$$

Notation: Let $a \in A$ denote an RGB tile (3 channels) and $b \in B$ a Sentinel-2 MSI tile (13 bands). $G : A \to B$ maps RGB→MSI and $F : B \to A$ maps MSI→RGB. D_A and D_B are least-squares GAN discriminators for domains A and B, respectively. $\mathbb{E}$ denotes expectation over the empirical data distributions p_A and p_B. For any pixel/voxel spectral vectors $x, y \in \mathbb{R}^{13}$, $\langle x,y\rangle$ is the Euclidean inner product and $\|x\|_2$ the ℓ_2 norm. $\mathcal{L}_{\mathrm{SAM}}(x,y)$ is the Spectral Angle Mapper (in radians unless otherwise stated); smaller is better. For histogram loss, S^+ and S^- are the sets of scalar spectral samples (e.g., band values) drawn from the target and generated MSI, respectively; B is the number of histogram bins, bin_b is the b-th bin, and h_b^+, h_b^- are the corresponding normalized bin frequencies. $\mathbf{1}\{\cdot\}$ is the indicator function. $\mathcal{L}_{\mathrm{GAN}}$ is the LSGAN objective, $\mathcal{L}_{\mathrm{cyc}}$ is the cycle-consistency term (here using SAM in place of pixel L1/L2), and $\mathcal{L}_{\mathrm{hist,tot}}$ aggregates the histogram losses in both directions. The generator objective is $\mathcal{L}_G = \mathcal{L}_{\mathrm{GAN}} + \lambda_{\mathrm{cycle}}\mathcal{L}_{\mathrm{cyc}} + \lambda_{\mathrm{hist}}\mathcal{L}_{\mathrm{hist,tot}}$, with $\lambda_{\mathrm{cycle}} = 10$ and $\lambda_{\mathrm{hist}} = 0.05$. $\mathcal{L}_{D_A}$ and $\mathcal{L}_{D_B}$ are discriminator losses for domains A and B.

To assess the realism of the multispectral images generated, a PatchGAN-based [30] discriminator is utilized. This discriminator employs four convolutional layers with progressively increasing feature map depth and decreasing spatial resolution. To stabilize training, LeakyReLU activations and instance normalization are applied. The final layer outputs a scalar value indicating whether the input image is real or synthetic.

Moreover, since the images in the Eurosat dataset are 64 × 64. During inference, we tested our model on 512 × 512 images using a 64 × 64 sliding window.

In addition, a specialized preprocessing pipeline was implemented for Sentinel-2 multispectral images. The SentinelToTensor transformation converts raw multispectral arrays into PyTorch tensors with channel-first ordering. The SentinelResize transformation ensures uniform image dimensions using bilinear interpolation, facilitating consistent model training. These preprocessing steps, combined with the Cycle GAN architecture, optimize the model's ability to generate realistic multispectral images from RGB inputs.

V. EXPERIMENTS

A. TEXT QUALITY ASSESSMENT FOR PIPELINE INTEGRATION

We evaluated the effectiveness of our proposed methodology by conducting multiple experiments. We first assess the performance of the Qwen2-VL-2B-Instruct model in generating 675,993 textual captions for satellite images by comparing the generated captions with the original captions from the SkyScript dataset. In Table 3, we reproduce the mean recall results reported by Wang et al. [28] and add our zero-shot evaluation of Qwen2-VL-2B-Instruct; under the same evaluation protocol, Qwen2-VL-2B-Instruct attains the highest mean recall among the compared models. The captions are measured as presented in Table 4.

TABLE 3. img2txt Mean recall (%) table for Skyscript Dataset.

Model	Mean recall
CLIP-original	2.97
Human-curated captions	3.28
CLIP-laion-RS	3.85
RemoteCLIP	5.08
SkyCLIP-30	8.53
Qwen2-VL-2B-Instruct	**23.36**

TABLE 4. Qwen2-VL-2B-Instruct caption similarities.

Metric	Value
BLEU-1	0.1567
BLEU-2	0.0526
BLEU-3	0.0298
BLEU-4	0.0183
METEOR	0.1353
ROUGE-L	0.1812
CIDEr-D	0.0132
BERT-F1	0.4636
CLIP Score	0.6822
RemoteCLIP	0.6569
SkyCLIP Score	0.7312

Overall, the results indicate that Qwen2-VL-2B-Instruct captures visual content reasonably well, but shows limited consistency with human-authored captions. The low scores for BLEU, METEOR, ROUGE-L and CIDEr-D reflect modest lexical and structural overlap, while the moderate BERT-F1 score suggests some degree of preserved semantic meaning. In particular, the relatively high CLIPScore (0.6822) demonstrates that the generated captions align well with the image content, even if they lack a longer coherent phrasing. This balance of strong visual grounding but weak phrase level detail supports the model's suitability for our text-to-image generation pipeline, while also highlighting room for improvement in linguistic fluency and contextual richness.

B. IMAGE QUALITY ASSESSMENT FOR PIPELINE INTEGRATION

To ensure the suitability of the fine-tuned Stable Diffusion 2-1 model within our pipeline, we evaluated the similarity between generated RGB images and real satellite images from the SkyScript dataset. As shown in Table 5, the model achieved a CMMD of 0.245, an FID of 31.5, and a KID of 0.0179 ± 0.0010. These values indicate a moderate resemblance between the generated and real image distributions, suggesting that the generated images are visually coherent and appropriate as intermediate data for our unified framework.

We also examined semantic alignment using different CLIP Scores, obtaining a value of 0.7312 (Table 4) between generated captions and generated images. This observation highlights the importance of the subsequent RGB-to-Sentinel-2 translation step, which enhances the spectral and contextual consistency of the final triplets. In contrast, this relatively high score suggests that the Qwen2-VL-2B-Instruct model successfully generates textual descriptions that are semantically similar to the original captions. Some examples of image and caption are presented in Fig. 3.

TABLE 5. Similarities of original and generated images.

Metric	Value
CMMD (CLIP Maximum Mean Discrepancy)	0.2450
FID	31.5480
KID	0.0179 ± 0.0010

TABLE 6. Cycle GAN inference metrics on EuroSAT and our synthetic images.

Metric	EuroSAT Dataset	Synthetic Images
SAM (°)	13.9374	10.1647
SID	0.0892	0.0705
ERGAS	23.2317	22.9375
MAE	3.8822	3.8272
MSE	18.7114	18.1829
CMMD	0.5289	0.6122

VI. RESULTS

The proposed methodology demonstrates a versatile pipeline capable of generating and transforming different modalities of satellite imagery and text descriptions. Specifically, the pipeline enables three distinct applications:

1) Caption-to-RGB and Multispectral Image Generation
2) RGB Image-to-Caption and Multispectral Image Translation
3) Multispectral Image-to-RGB and Caption Generation

Each of these transformations is facilitated by the integration of fine-tuned Stable Diffusion and Cycle GAN models, enabling a bidirectional relationship between textual descriptions, RGB imagery, and multispectral data.

When only a textual description (caption) is available, the fine-tuned Stable Diffusion 2-1 Base model is capable of generating a realistic synthetic RGB satellite image based on the input caption. The generated image maintains key structural elements described in the text, including features such as water bodies, vegetation, roads, and urban areas. However, because of the limitations of diffusion models, some fine-grained details and spectral characteristics may not be perfectly aligned with real-world satellite images.

Once the synthetic RGB image is generated, it is passed through the Cycle GAN model, which translates it into a 13-band Sentinel-2 multispectral image. The Cycle GAN model has been trained to map RGB textures to the corresponding spectral responses, ensuring that the resulting multispectral image retains realistic spectral information. This approach provides a way to generate plausible multispectral satellite data from captions, which can be useful in scenarios where real multispectral observations are unavailable or incomplete (see Figure 4). When an RGB satellite image is available without any associated metadata, the proposed pipeline can generate a corresponding textual caption and a multispectral version of the image. Notably, while metrics

FIGURE 3. Examples of original and generated image–caption pairs. Each panel presents an original SkyScript image with its reference caption (top) alongside a generated image produced by the fine-tuned Stable Diffusion model conditioned on captions generated by Qwen2-VL-2B-Instruct (bottom). The generated captions are generally, which often contain rigid land-use terms. The generated images capture overall scene characteristics but occasionally differ in fine details from the originals. These examples illustrate both the potential and the limitations of our pipeline in producing semantically coherent image–caption pairs for remote sensing.

improve on our synthetic set, reflecting a larger distributional gap in high-level semantics: Stable Diffusion images are smoother and less noisy [17], [49] making RGB→MSI mapping easier yet they diverge more from real EuroSAT scenes in CLIP feature space.

Figure 5 presents the loss curves during training and validation. During inference, the performance of multispectral conversion was evaluated using several quantitative metrics. Testing was conducted on 512 × 512 images by randomly cropping a 64 × 64 patch from each image in our stable diffusion generated dataset and Eurosat dataset which is in 64 × 64 size. The following table (Table 6) summarizes the results along with the corresponding references for each metric.

VII. DISCUSSION

As a measure of performance, we calculated some key metrics that capture different aspects of image quality. The Spectral Angle Mapper (SAM) is an estimate of the angular difference between the spectral vectors of the target and the generated images. The SAM on Eurosat is 13.9374°, whereas it was 10.1647° in our generated dataset, so that the spectral fidelity is better preserved in the latter. Spectral Information Divergence (SID) measures the extent to which probability distributions of the spectral signature are deviating. Eurosat generated an SID of 0.0892, while our synthesized dataset achieved 0.0705. Lower SID guarantees greater spectral consistency between generated and reference images. ERGAS (relative global error) condenses the overall reconstruction error. In the Eurosat test images, ERGAS was 23.2317; in our dataset, it dropped to 22.9375, a measure of a moderate increase in global precision. The mean Absolute Error (MAE) and the mean Squared Error (MSE) quantify the pixel-wise intensity changes. Since our evaluation is unpaired generated and reference images are not pixel-for-pixel identical, MAE and MSE readings are greater compared to paired situations. But MAE decreased from 3.8822 on Eurosat to 3.8272 on our images, and MSE fell from 18.7114 to 18.1829. These reductions indicate better per-pixel consistency despite the fact that the test is unpaired.

The CLIP maximum mean discrepancy (CMMD) was 0.5289 on Eurosat but increased to 0.6122 on our synthesized set, which shows slightly more variability in the distributions of higher-order moments. Although our Stable

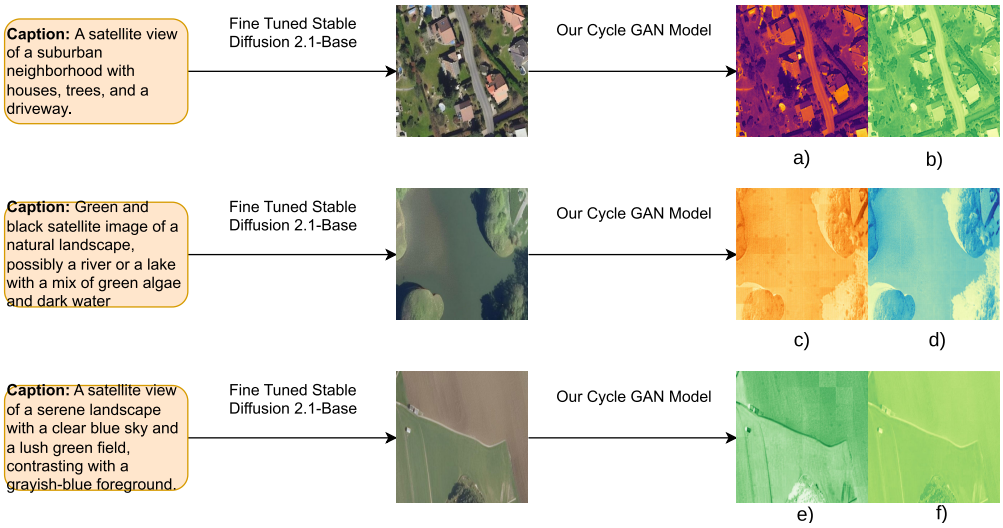

FIGURE 4. Text-to-spectrum generation pipeline. From left to right: natural-language captions; RGB tiles synthesized by our fine-tuned Stable Diffusion 2.1-Base; and outputs from our CycleGAN that translate RGB into spectral products.(a) Thermal Spectrum, (b) NDVI Spectrum, (c) Short Wave Infrared (SWIR) Spectrum, (d) Bathymetric Spectrum, (e) Agriculture Spectrum, (f) NDVI Spectrum.

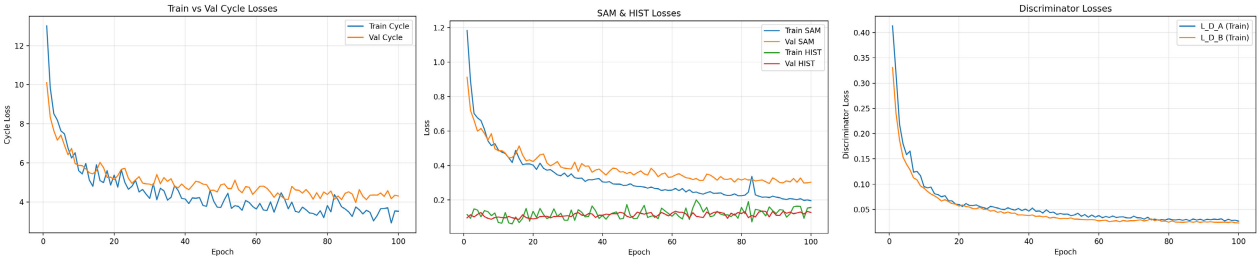

FIGURE 5. Left Graphics: Generator Loss, Middle Graphics: SAM and Histogram Losses, Right Graphics: Discriminator Losses during training.

Diffusion images are highly diverse, they inherently lack the real-world sensor noise, atmospheric distortions, and fine spectral variability present in held-out Eurosat tiles, so Cycle GAN's learned mapping reproduces these synthetic scenes with lower SAM, SID, ERGAS, MAE, and MSE, yielding seemingly better metrics than on actual Eurosat data. Beyond evaluation scores, Fig.6, 7 provides a band-wise qualitative check showing the patterns, across all 13 Sentinel-2 bands. As a result, Cycle GAN preserves pixel-wise spectral consistency and global distribution of spectral signatures. We recommend the 64×64 sliding window approach for inference in higher resolution images (see Figure 6).

Although a full benchmark is out of scope, the released triplets are directly usable in common RS applications: for land cover classification, they enable simple linear-probe tests with triplet-based augmentation; for geospatial retrieval, text→RGB search can be followed by MSI re-ranking while jointly reporting recall@k and spectral consistency (e.g., SAM); and for environmental monitoring, text-conditioned pairing of pre/post tiles allows computing MSI change proxies (e.g., NDVI deltas). We provide minimal scripts to instantiate these protocols, so semantic relevance and spectral fidelity can be measured side by side.

Our multispectral products are learned translations from RGB and are therefore not radiometric ground truth. They may differ from real Sentinel-2 in absolute reflectance, bandpass/MTF characteristics, georegistration, and noise/atmospheric artifacts. Consequently, while useful for semantic tasks and prototyping, the synthetic MSI should not be used as a substitute for calibrated Level-2 surface reflectance or physics-based retrieval.

The pipeline adapts by swapping the target band set and sensor priors and retraining the RGB→MSI translator. In practice, outputs are remapped to the new sensor's spectral response functions, the translator is optionally conditioned on a sensor identifier to support multiple sensors. To address ethical considerations and potential misuse, all releases include clear disclaimers that outputs are synthetic (not radiometric ground truth) and are governed by explicit safeguards-Apache-2.0 for code, CC BY-NC 4.0 with an Acceptable Use Policy for data, and a Responsible AI Model License for weights that prohibit operational/surveillance/safety critical

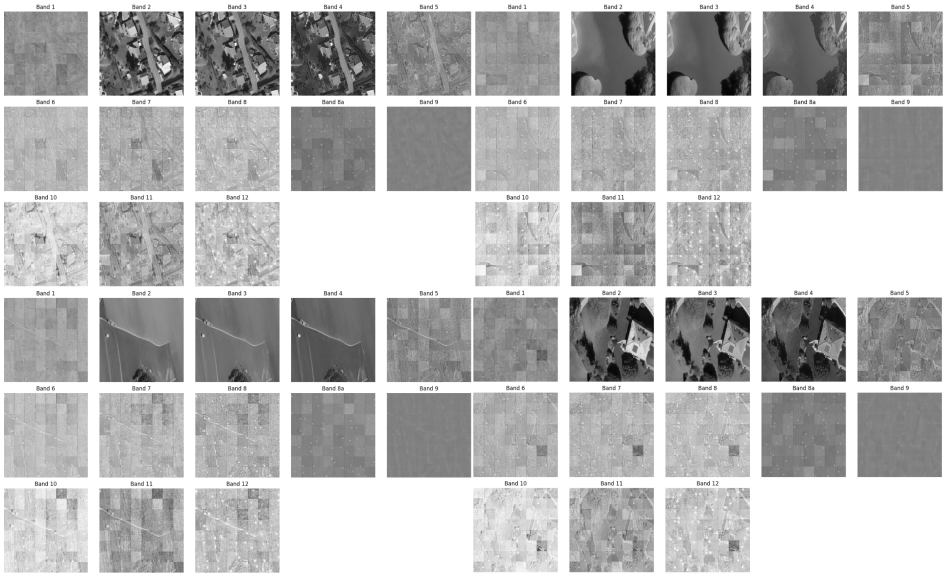

FIGURE 6. Illustration of the all Sentinel-2 bands for some example images.

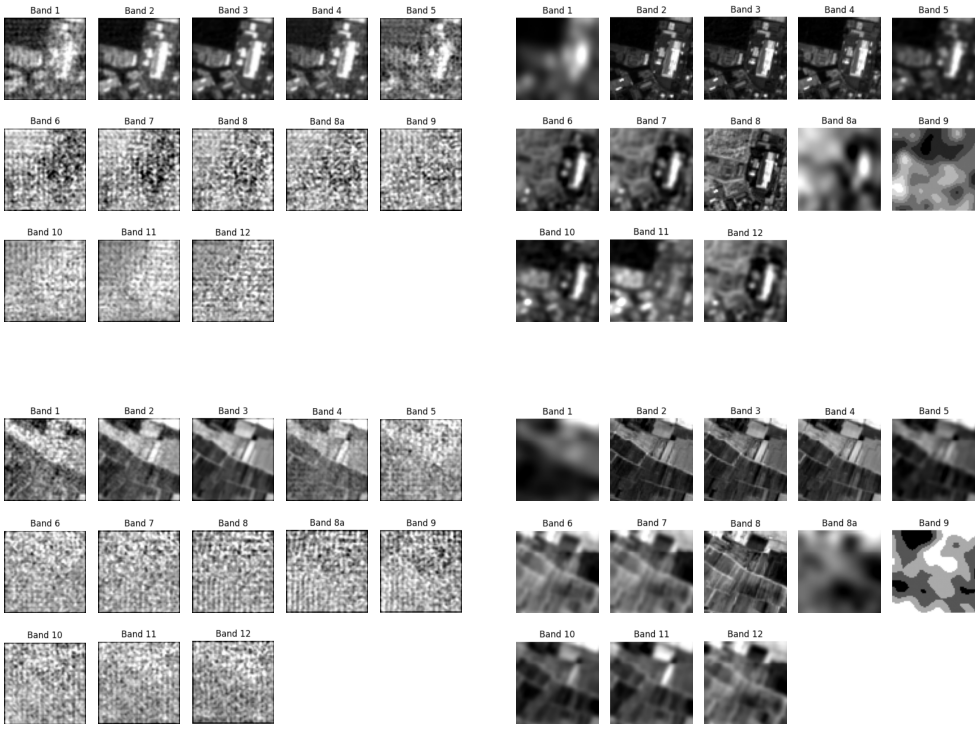

FIGURE 7. Illustration of the all Sentinel-2 bands for some examples in the EuroSAT Dataset. Left side: Generated images, Right side: Original images.

use, require transparent disclosure, and mandate provenance tracking.

Our study has several limitations. First, computational and memory constraints (single-GPU training; peak $\approx$ 22.5 GB VRAM) forced the use of small patch sizes (64 $\times$ 64 for training and sliding-window inference at 512 $\times$ 512) and prevented a thorough comparison against alternative models (e.g., Pix2Pix / TransCycleGAN) or larger ablation studies. Second, we evaluated our synthetic dataset against the EuroSAT dataset, so pixel-wise metrics (e.g., MAE/MSE) do not reflect exact correspondence and can misestimate fidelity; likewise, while SAM/SID/ERGAS summarize spectral behavior, they do not constitute full radiometric validation. Third, the text-to-image stage shows imperfect

alignment (e.g., lower caption–image CLIP alignment than caption–caption), which may propagate semantic mismatches into the multispectral translation. Fourth, training and testing were limited to the EuroSAT distribution and Sentinel-2-like targets; generalization across sensors, resolutions, geographies, and seasons remains unverified, and our synthetic images lack real-world sensor noise and atmospheric effects. Moreover, our pipeline composes captioning, text-to-image synthesis, and RGB-to-multispectral translation, so errors may accumulate across stages. In particular, underspecified or inaccurate captions can induce semantic drift in the synthetic RGB (reflected in lower image–text alignment), which then propagates as spectral discrepancies during RGB→MSI mapping. To the best of our knowledge, there is no public end-to-end caption→image→multispectral pipeline with shared code and standardized metrics, which precludes a direct system-level comparison. Moreover, our pipeline composes captioning, text-to-image synthesis, and RGB-to-multispectral translation, so errors may accumulate across stages. In particular, underspecified or inaccurate captions can induce semantic drift in the synthetic RGB (reflected in lower image–text alignment), which then propagates as spectral discrepancies during RGB→MSI mapping. Finally, natural-language captions may not describe of complex scenes which are highly variable, fine-grained spatial patterns well. Future work includes scaling compute to enable broader baselines and ablations, and exploring stronger text–image alignment under realistic acquisition conditions.

VIII. CONCLUSION

In this study, we proposed a multimodal pipeline that integrates text-to-image generation, RGB-to-multispectral conversion, and captioning models for synthetic satellite image generation and transformation. Our approach enables three key functionalities: generating RGB and multispectral images from textual descriptions, deriving captions and multispectral data from RGB images, and reconstructing RGB and captions from multispectral images. Experimental results demonstrate that the proposed method generates visually plausible satellite images while preserving spectral consistency. However, challenges remain in text-to-image alignment, multispectral conversion accuracy, and caption fluency.

ACKNOWLEDGMENT

The authors are grateful for the HPC resources provided by the IT Research Center of Vilnius University.

APPENDIX

See Figs 6 and 7

REFERENCES

[1] V. V. Cepeda, G. K. Nayak, and M. Shah, "GeoCLIP: Clip-inspired alignment between locations and images for effective worldwide geo-localization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 8690–8701.

[2] Z. Yan, J. Li, X. Li, R. Zhou, W. Zhang, Y. Feng, W. Diao, K. Fu, and X. Sun, "RingMo-SAM: A foundation model for segment anything in multimodal remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5625716.

[3] Y. Zhang, M. Ye, G. Zhu, Y. Liu, P. Guo, and J. Yan, "FFCA-YOLO for small object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5611215.

[4] D. Szwarcman et al., "Prithvi-EO-2.0: A versatile multi-temporal foundation model for earth observation applications," 2024, *arXiv:2412.02732*.

[5] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, A. Meygret, F. Spoto, O. Sy, F. Marchese, and P. Bargellini, "Sentinel-2: ESA's optical high-resolution mission for GMES operational services," *Remote Sens. Environ.*, vol. 120, pp. 25–36, May 2012.

[6] S. Lu, J. Guo, J. R. Zimmer-Dauphinee, J. M. Nieusma, X. Wang, P. VanValkenburgh, S. A. Wernke, and Y. Huo, "AI foundation models in remote sensing: A survey," 2024, *arXiv:2408.03464*.

[7] L. Zhang and L. Zhang, "Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 2, pp. 270–294, Jun. 2022.

[8] P. F. Rozario, J. Lee, Y. Chen, P. D. Mohan, M. DeWitte, and R. Gomes, "Analyzing the impact of geospatial derivatives on domain adaptation with CycleGAN," in *Proc. IEEE Int. Conf. Electro Inf. Technol. (eIT)*, May 2024, pp. 710–715.

[9] L. Liu, B. Chen, H. Chen, Z. Zou, and Z. Shi, "Diverse hyperspectral remote sensing image synthesis with diffusion models," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5532616.

[10] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[11] B. Liu, J. Ji, L. Gu, and Z. Jiang, "An integrated CycleGAN-diffusion approach for realistic image generation," *Proc. SPIE*, vol. 29, pp. 91–101, Apr. 2024.

[12] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, Apr. 2018.

[13] Q. Cheng, H. Huang, Y. Xu, Y. Zhou, H. Li, and Z. Wang, "NWPU-captions dataset and MLCA-net for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5629419.

[14] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, Jul. 2019.

[15] Z. Yuan, Z. Xiong, L. Mou, and X. X. Zhu, "ChatEarthNet: A global-scale image–text dataset empowering vision–language geo-foundation models," *Earth Syst. Sci. Data*, vol. 17, no. 3, pp. 1245–1263, Mar. 2025.

[16] Z. Wang, R. Prabha, T. Huang, J. Wu, and R. Rajagopal, "SkyScript: A large and semantically diverse vision-language dataset for remote sensing," in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2024, pp. 5805–5813.

[17] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6840–6851.

[18] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–18.

[19] Q. Liu, H. Zhou, Q. Xu, X. Liu, and Y. Wang, "PSGAN: A generative adversarial network for remote sensing image pan-sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10227–10242, Dec. 2021.

[20] Y. Xiao, Q. Yuan, K. Jiang, J. He, X. Jin, and L. Zhang, "EDiffSR: An efficient diffusion probabilistic model for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5601514.

[21] C. X. Ren, A. Ziemann, J. Theiler, and A. M. S. Durieux, "Cycle-consistent adversarial networks for realistic pervasive change generation in remote sensing imagery," in *Proc. IEEE Southwest Symp. Image Anal. Interpretation (SSIAI)*, Mar. 2020, pp. 42–45.

[22] L. Abady, J. Horváth, B. Tondi, E. J. Delp, and M. Barni, "Manipulation and generation of synthetic satellite images using deep learning models," *J. Appl. Remote Sens.*, vol. 16, no. 4, Nov. 2022, Art. no. 046504.

[23] H. Tang, H. Liu, D. Xu, P. H. S. Torr, and N. Sebe, "AttentionGAN: Unpaired image-to-image translation using attention-guided generative adversarial networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 4, pp. 1972–1987, Apr. 2023.

[24] J. D. Bermudez, P. N. Happ, R. Q. Feitosa, and D. A. B. Oliveira, "Synthesis of multispectral optical images from SAR/optical multitemporal data using conditional generative adversarial networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1220–1224, Aug. 2019.
[25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn.*, 2022, pp. 8748–8763.
[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[27] K. Kuckreja, M. S. Danish, M. Naseer, A. Das, S. Khan, and F. S. Khan, "GeoChat: Grounded large vision-language model for remote sensing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 27831–27840.
[28] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin, "Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution," 2024, *arXiv:2409.12191*.
[29] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10674–10685.
[30] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.
[31] Z. Zhang, T. Zhao, Y. Guo, and J. Yin, "RS5M and GeoRSCLIP: A large scale vision-language dataset and a large vision-language model for remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2023, Art. no. 5642123.
[32] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 204–207.
[33] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.
[34] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, 2005, pp. 65–72.
[35] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Workshop Text Summarization Branches Out*, 2004, pp. 74–81.
[36] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4566–4575.
[37] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with BERT," 2019, *arXiv:1904.09675*.
[38] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 7514–7528.
[39] F. Liu, D. Chen, Z. Guan, X. Zhou, J. Zhu, Q. Ye, L. Fu, and J. Zhou, "RemoteCLIP: A vision language foundation model for remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5622216.
[40] Z. Liang, Y. Xu, Y. Hong, P. Shang, Q. Wang, Q. Fu, and K. Liu, "A survey of multimodel large language models," in *Proc. 3rd Int. Conf. Comput., Artif. Intell. Control Eng.*, Jan. 2024, pp. 405–409.
[41] S. Sastry, S. Khanal, A. Dhakal, and N. Jacobs, "GeoSynth: Contextually-aware high-resolution satellite image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2024, pp. 460–470.
[42] J. Ma, J. Jiang, A. Fan, J. Jiang, and J. Yan, "Remote sensing image registration: A comprehensive survey," *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 166–188, Jun. 2019.
[43] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.
[44] S. P. Tadem, "Cyclegan with three different unpaired datasets," 2022, *arXiv:2208.06526*.
[45] A. Bourou, K. Daupin, V. Dubreuil, A. De Thonel, V. Mezger-Lallemand, and A. Genovesio, "Unpaired image-to-image translation with limited data to reveal subtle phenotypes," in *Proc. IEEE 20th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2023, pp. 1–5.
[46] Y. Zi, F. Xie, X. Song, Z. Jiang, and H. Zhang, "Thin cloud removal for remote sensing images using a physical-model-based CycleGAN with unpaired data," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 1004605.
[47] R. H. Yuhas, A. Goetz, and J. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm," in *Proc. 3rd Summaries Annu. JPL Airborne Geosci. Workshop*, vol. 1, 1992, pp. 1–16.
[48] E. Ustinova and V. Lempitsky, "Learning deep embeddings with histogram loss," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2022, pp. 4170–4178.
[49] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *Proc. Special Interest Group Comput. Graph. Interact. Techn. Conf.*, Aug. 2022, pp. 1–10.

KURSAT KOMURCU received the B.Sc. degree in electronics and communication engineering from Yildiz Technical University, Esenler, Türkiye, in 2023, and the M.S. degree in informatics from the Institute of Computer Science, Vilnius University, Vilnius, Lithuania, in 2025. His research interests are focused on computer vision and deep learning.

LINAS PETKEVICIUS (Member, IEEE) received the Ph.D. degree in informatics from the Institute of Computer Science, Vilnius University, Vilnius, Lithuania, in 2020. Since 2022, he has been the Head of the Software Engineering Department, Institute of Computer Science. His research interests are focused on computer vision and deep learning, as well as statistical inference and outlier detection.