Anomaly Detection in Aerial Videos With Transformers

Pu Jin, *Member, IEEE*, Lichao Mou, Gui-Song Xia, *Senior Member, IEEE*, and Xiao Xiang Zhu, *Fellow, IEEE*

***Abstract*— Unmanned aerial vehicles (UAVs) are widely applied for purposes of inspection, search, and rescue operations by the virtue of low-cost, large-coverage, real-time, and high-resolution data acquisition capacities. Massive volumes of aerial videos are produced in these processes, in which normal events often account for an overwhelming proportion. It is extremely difficult to localize and extract abnormal events containing potentially valuable information from long video streams manually. Therefore, we are dedicated to developing anomaly detection methods to solve this issue. In this article, we create a new dataset, named Drone-Anomaly, for anomaly detection in aerial videos. This dataset provides 37 training video sequences and 22 testing video sequences from seven different realistic scenes with various anomalous events. There are 87 488 color video frames (51 635 for training and 35 853 for testing) with the size of 640 × 640 at 30 frames/s. Based on this dataset, we evaluate existing methods and offer a benchmark for this task. Furthermore, we present a new baseline model, anomaly detection with Transformers (ANDTs), which treats consecutive video frames as a sequence of tubelets, utilizes a Transformer encoder to learn feature representations from the sequence, and leverages a decoder to predict the next frame. Our network models normality in the training phase and identifies an event with unpredictable temporal dynamics as an anomaly in the test phase. Moreover, to comprehensively evaluate the performance of our proposed method, we use not only our Drone-Anomaly dataset but also another dataset. We will make our dataset and code publicly available. A demo video is available at https://youtu.be/ancczYryOBY. We make our dataset and code publicly available (https://gitlab.lrz.de/ai4eo/reasoning/drone-anomaly https://github.com/Jin-Pu/Drone-Anomaly).**

***Index Terms*— Aerial videos, anomaly detection, convolutional neural networks (CNNs), temporal reasoning, transformers, unmanned aerial vehicle (UAV).**

Manuscript received 6 December 2021; revised 30 March 2022 and 27 June 2022; accepted 7 July 2022. Date of publication 11 August 2022; date of current version 29 August 2022. This work was supported in part by the European Research Council (ERC) through the European Union's Horizon 2020 Research and Innovation Program (*So2Sat*) under Grant ERC-2016-StG-714087, in part by the Helmholtz Association through the Framework of Helmholtz Artificial Intelligence (AI) under Grant ZT-I-PF-5-01, in part by the Local Unit "Munich Unit@Aeronautics, Space and Transport (MASTr)" and Helmholtz Excellent Professorship "Data Science in Earth Observation—Big Data Fusion for Urban Research" under Grant W2-W3-100, in part by the German Federal Ministry of Education and Research (BMBF) through the Framework of the International Future AI Lab—"Artificial Intelligence for Earth Observation (AI4EO): Reasoning, Uncertainties, Ethics and Beyond" under Grant 01DD20001, and in part by the German Federal Ministry for Economic Affairs and Climate Action through the Framework of the "National Center of Excellence ML4Earth" under Grant 50EE2201C. *(Corresponding authors: Lichao Mou; Xiao Xiang Zhu.)*

Pu Jin, Lichao Mou, and Xiao Xiang Zhu are with the Remote Sensing Technology Institute, German Aerospace Center, 82234 Weßling, Germany, and also with the Data Science in Earth Observation (former: Signal Processing in Earth Observation), Technical University of Munich, 80333 Munich, Germany (e-mail: pu.jin@dlr.de; lichao.mou@dlr.de; xiaoxiang.zhu@dlr.de).

Gui-Song Xia is with the State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS) and the School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: guisong.xia@whu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2022.3198130

I. Introduction

ANOMALY detection refers to the detection of visual instances that significantly deviate from the majority [1]. Due to the expanding demand in broad domains, such as inspection [2], [3], [4], [5], [6], search operations [7], [8], and security [9], [10], [11], [12], anomaly detection plays increasingly important roles in various communities, including computer vision, data mining, machine learning, and remote sensing. With the proliferation of unmanned aerial vehicles (UAVs) worldwide, massive produced aerial videos spur the demand for detecting abnormal events in aerial video sequences in a wide range of applications [13]. For example, many long-endurance UAVs[1] are developed and utilized in inspection operations [2], [3], [4], [5], [6]. Large amounts of aerial videos are created by these UAVs, in which normal video segments often account for an overwhelming proportion of the whole video. It is time-consuming and costly to find potentially valuable information from long and untrimmed videos manually. Therefore, we are intended to adopt anomaly detection methods to temporally localize anomalous events in aerial videos automatically.

Usually, we cannot know beforehand what anomalies are in a scene, because there are too many possibilities that are impossible to be exhaustively listed. By contrast, it is easy to have information on the nature of normality in advance. Hence, most existing methods for anomaly detection only use normal data to learn feature representations of normality and consider test instances that cannot be well described as anomalies. Massive studies [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26] are dedicated to detecting and categorizing non-conforming patterns present in images. These studies mainly focus on spatial occurrences of anomalous patterns. In contrast, anomaly detection in videos aims at identifying temporal occurrences (i.e., start and end times) of abnormal events. In computer vision, many methods [27], [28], [29], [30], [31], [32], [33], [34] have been

[1] https://www.airforce-technology.com/features/featurethe-top-10-longest-range-unmanned-aerial-vehicles-uavs/

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

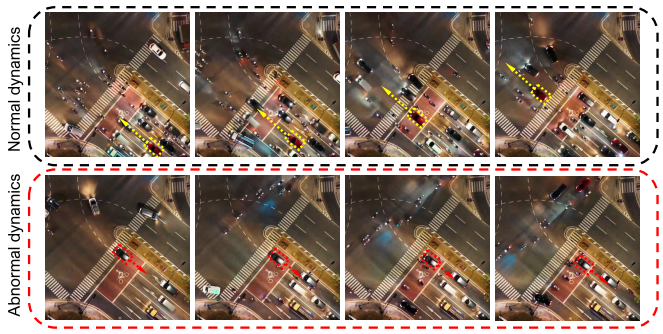


Fig. 1. Abnormal and normal dynamics. We display some frames from the *crossroads* scene for demonstrating the importance of temporal information in detecting anomalous events in aerial videos. In the normal video clip (top), all vehicles have a consistent moving direction. We use a yellow box with an arrow to represent an example vehicle and its moving direction. In the abnormal video snippet (bottom), a vehicle (in the red box) moves backward on the road. We can see the importance of temporal context in this task.

proposed for this task in surveillance videos. In comparison with surveillance videos, UAV videos bring the following challenges: 1) moving camera instead of static camera and 2) variable spatial resolution due to changes in flight altitude. Existing works [35], [36], [37] predefine several categories of anomalous events, convert aerial video anomaly detection into an event recognition task, and utilize supervised methods to address this problem. By contrast, in this work, we are interested in unsupervised methodologies for this task. Because in many real-world applications, it is not possible to exhaustively list all anomalous events beforehand. More specifically, we train a model for anomaly detection in aerial videos using only normal data that can be collected easily in advance.

In this article, we focus on detecting anomalous events in aerial videos. To this end, we create a new dataset, named Drone-Anomaly, providing 37 training video sequences and 22 testing video sequences from seven different realistic scenes. The dataset contains real-world anomalous events that are not staged by actors. Based on this dataset, we evaluate existing methods and offer a benchmark. In addition, we note that the modeling temporal context is critical (see Fig. 1). Most existing anomaly detection methods utilize convolution-based encoders for capturing spatiotemporal dependencies among the input video frames. However, this is limited in learning long-term relations due to limited temporal receptive fields of these models. In this article, we present a new baseline model, anomaly detection with Transformers (ANDTs), which takes as input several consecutive video frames, leverages a Transformer encoder to model global context, and utilizes a decoder to predict the next frame. More specifically, ANDT treats a video as a sequence of tubelets and maps them into tubelet embeddings by linear projection. For preserving spatiotemporal information, the tubelet embeddings are added with learnable spatiotemporal position embeddings and then fed into a Transformer encoder to learn a spatiotemporal feature. The decoder is subsequently combined with the encoder for predicting the next frame based on the learned spatiotemporal representation. Our network is able to well predict an event with normal temporal dynamics and identifies an event with unpredictable temporal dynamics as an anomaly in the test phase.

The main contributions of this article can be summarized as follows.

1) We create an annotated dataset consisting of 37 training videos and 22 testing videos involving seven realistic scenes, covering a large variety of anomalous events. This dataset expands the scope of anomaly detection research. In addition, we extensively validate existing methods to provide a benchmark for this task.
2) We extensively validate existing methods to provide a challenging benchmark for anomaly detection in aerial videos.
3) We present a new baseline model ANDT and conduct extensive ablation studies and experiments for validating the effectiveness of our approach. To the best of our knowledge, this is the first time that a Transformer-based network is proposed for video anomaly detection.

The remaining sections of this article are organized as follows. The related works are introduced in Section II. Then, we detail our new dataset in Section V-A. Also, our network is described in Section IV. Section V shows and discusses experimental results. Finally, this article is concluded in Section VI.

II. Related Work

In remote sensing, there have been a number of works for anomaly detection in hyperspectral imagery [40], [41], [42], [43], [44], [45], [46], [47], [48]. These studies mainly focus on locating pixels with significantly different spectral signatures from their neighboring background pixels in the spatial domain. For example, the Reed-Xiaoli (RX) algorithm [40] uses a local Gaussian model to detect anomalies in hyperspectral images and has become a baseline model. In [41], a collaborative representation detector (CRD) is proposed to detect pixels with unknown spectral signatures. Recently, deep learning-based methods have drawn significant attention. Chen *et al.* [42] propose to use an autoencoder to learn representative features to detect anomalies in an unsupervised manner. Hu *et al.* [43] employ convolutional neural networks (CNNs) to learn spectral–spatial features in this task and achieve outstanding performance.

From static imagery to multitemporal images, much effort [49], [50], [51], [52], [53], [54], [55], [56] has been made to detect anomalies in the temporal domain. For instance, [49] uses multispectral images over two years for locating and identifying crop anomalies in two soybean fields. Liu *et al.* [50] leverage multitemporal thermal infrared (TIR) images for detecting geothermal anomaly areas by spatiotemporal analysis. In [51], multitemporal Landsat images are utilized to detect normalized difference vegetation index (NDVI) anomalies for mapping incongruous patches in coffee plantations.

Moreover, we notice that in computer vision, many anomaly detection approaches [57], [58], [59], [60], [61], [62], [63] have been developed for fixed camera surveillance videos. By contrast, we think that anomaly detection in aerial videos is more challenging, because the videos are usually acquired

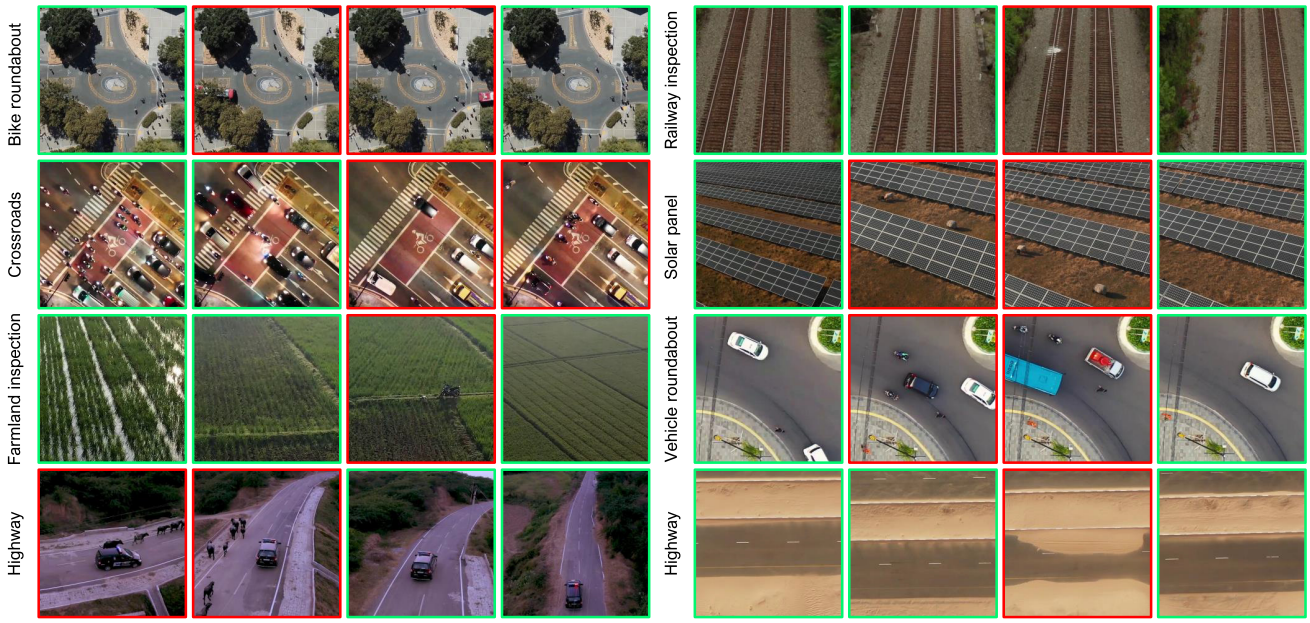


Fig. 2. Overview of the Drone-Anomaly dataset. We show four frames of each video. The anomalous frames are marked with red borders, and frames with green borders are normal ones.

by moving cameras. There have been a few works for investigating anomaly detection in aerial videos. These works [35], [36], [37] regard this problem as an event recognition task. Specifically, they first predefine several anomalous activities and then leverage supervised methods to recognize the defined events from aerial videos. For example, [35] leverages object tracking and classification methods to obtain trajectories and semantic information and then utilizes an ontology-based reasoning model to learn spatiotemporal relations among them for detecting video events. Yang *et al.* [36] define three different safety-related anomalies and propose a functional approach that models temporal relations of time-to-collision safety indicators to detect these anomalies from UAV-based traffic videos. Furthermore, [37] proposes a hybrid approach that integrates trajectories and semantic information of objects to build high-level knowledge for extracting complicated critical activities and events from UAV videos. Most recently, based on the AU-AIR dataset [64] that is proposed for object detection in UAV videos, [39] builds a dataset including several anomalous objects (hereafter, we call it AU-AIR-Anomaly dataset) and proposes a supervised method, a deep neural network-based context-aware anomaly detection method (CADNet), to detect instances and contextual anomalies in aerial videos. Compared with our dataset, the AU-AIR-Anomaly dataset only contains a single scene, i.e., traffic, and its aerial video has a relatively stable perspective.

In real-world applications, there are many possible anomalies existing in a scenario, which cannot be exhaustively listed and defined in advance. Instead, the nature of normality is relatively stable and easy to know beforehand. Therefore, we propose an unsupervised method ANDT that learns feature representations of genetic normality from merely normal data and determines test data with large reconstruction errors as anomalies. Moreover, methods [35], [36], [37], [39], [65] all leverage convolution-based encoders for learning spatiotemporal dependencies among input video frames. Due to the limited temporal receptive fields, these models are unable to effectively capture long-term temporal relations. By contrast, our method ANDT adopts a Transformer-based encoder that confers our model with a global temporal receptive field and enables it to capture temporal dependencies among all input frames. With a global perspective, our model is adept at distinguishing the movement of instances from the dynamic background and provides rich contextual information for detecting anomalies.

III. Dataset

To address the lack of available datasets for anomaly detection in aerial videos, we present the Drone-Anomaly. This section introduces the construction of our dataset, including video collection and annotation. Finally, we present the overall statistics of the dataset.

A. Video Collection

We collect aerial videos on YouTube[2] and Pexels[3] using search queries (e.g., *drone highway* and *UAV roundabout*) for each scene. To increase the diversity of anomalous events, we retrieve aerial videos using different languages (e.g., English, Chinese, German, and French). Moreover, to ensure the quality of aerial videos, we remove videos with any of the following situations: too short duration, manually edited, not captured by UAV cameras, and without clear anomalous events. We show four frames of an example video from each scene in Fig. 2.

[2]https://www.youtube.com/
[3]https://www.pexels.com/

TABLE I
DATASET DETAILS. WE PROVIDE VARIABLE DETAILS OF THE DRONE-ANOMALY DATASET

Scene	# Video snippets (Train / Test)	# Frames (Train / Test)	Example anomalies
Highway	6 / 3	9045 / 2820	Animals walking on the street; Car collision
Crossroads	10 / 5	15772 / 6244	Retrograde vehicles; Traffic congestion
Bike roundabout	6 / 7	7950 / 18427	Moving vehicles
Vehicle roundabout	4 / 2	5266 / 2643	People crossing the road
Railway inspection	3 / 1	1206 / 882	Obstacles on the railway
Solar panel inspection	4 / 3	2848 / 2450	Unknown objects; Defects of panel
Farmland inspection	4 / 1	9548 / 2387	Unidentified vehicles

TABLE II
COMPARISON WITH RELATED DATASETS. WE OFFER VARIOUS COMPARISIONS FOR EACH DATASETS

Dataset	# Videos	# Frames	# Scenes	Type of task	Type of anomalies	Year
Mini-drone [38]	38	22,860	1	Event recognition and detection	Actor-staged anomalies	2015
AU-AIR-Anomaly*[39]	1	32,823	1	Anomaly detection	Realistic anomalies	2021
Drone-Anomaly	59	87,488	7	Anomaly detection	Realistic anomalies	2022

* The AU-AIR dataset is originally created for object detection tasks.

B. Annotation

We assign video-level labels for training data. In the test phase, frame-level annotations are needed to evaluate the performance. Thus, we provide frame-level labels with binary values, where anomalous frames are labeled as 1, and 0 indicates normal frames. For each scene, training videos and testing videos with anomalies are provided. The details are shown in Table I.

C. Statistics

Our Drone-Anomaly dataset consists of long, untrimmed aerial videos that cover seven real-world scenes, including *highway*, *crossroads*, *bike roundabout*, *vehicle roundabout*, *railway inspection*, *solar panel inspection*, and *farmland inspection*. Various anomalies in these scenes have important practical significance and applications. We provide the overview of our dataset in Table I. Basically, the dataset consists of 37 training video sequences and 22 testing sequences. Each of them is at 30 frames/s and with a spatial size of 640×640 pixels. There are a total of 87 488 color video frames (51 635 for training and 35 853 for testing).

D. Comparison With Related Datasets

We compare our dataset with related datasets in Table II. Mini-drone dataset [38] consisting of 38 videos is proposed to parse video contents for privacy protection. The dataset contains three categories: normal, suspicious, and illicit behaviors. All events are staged by actors. This dataset can be used for different tasks, e.g., action recognition, video classification, event recognition, and event detection. In addition, based on the AU-AIR dataset [64], [39] annotates different anomalous events for detecting anomalies in aerial videos. The AU-AIR-Anomaly dataset contains four realistic anomalies, i.e., *a car on a bike road*, *a person on a road*, *a parked van in front of a building*, and *a bicycle on a road*.

IV. METHODOLOGY

In this section, we detail our model. First, we introduce future frame prediction—the framework we use for anomaly detection, in Section IV-A. Next, we give the detailed description of ANDT in Section IV-B.

A. Future Frame Prediction for Anomaly Detection

For anomaly detection in aerial videos, comparing with the commonly used reconstruction-based framework [31], [66], [67], [68], [69], [70], [71], [72], [73], [74] where target values are equal to the inputs, it is more natural to predict the next video frame conditioned on several consecutive frames and compare the predicted one with its ground truth. In this way, temporal context can be modeled. The assumption of the future frame prediction framework is that temporal consistency in normal events is maintained stably; thus, normal events are temporally more predictable than anomalies. In the training stage, a network is trained with only normal videos to learn normal temporal patterns. In the test phase, events and activities not perfectly predicted by the network are then deemed as anomalies. Formally, given a video $\mathcal{V}$ composed of consecutive T frames, $\mathcal{V} = \{\boldsymbol{I}_1, \boldsymbol{I}_2, \ldots, \boldsymbol{I}_T\}$. All frames are stacked temporally and then utilized to predict the next frame $\boldsymbol{I}_{T+1}$. The predicted frame is denoted as $\hat{\boldsymbol{I}}_{T+1}$. We aim to learn a mapping $\mathcal{P}$ as follows:

$$\mathcal{P}(\mathcal{V}) \rightarrow \hat{\boldsymbol{I}}_{T+1}. \tag{1}$$

To make $\hat{\boldsymbol{I}}_{T+1}$ closer to $\boldsymbol{I}_{T+1}$, we minimize their ℓ_2 distance in intensity space as follows:

$$L\left(\hat{\boldsymbol{I}}, \boldsymbol{I}\right) = \left\| \hat{\boldsymbol{I}} - \boldsymbol{I} \right\|_2^2. \tag{2}$$

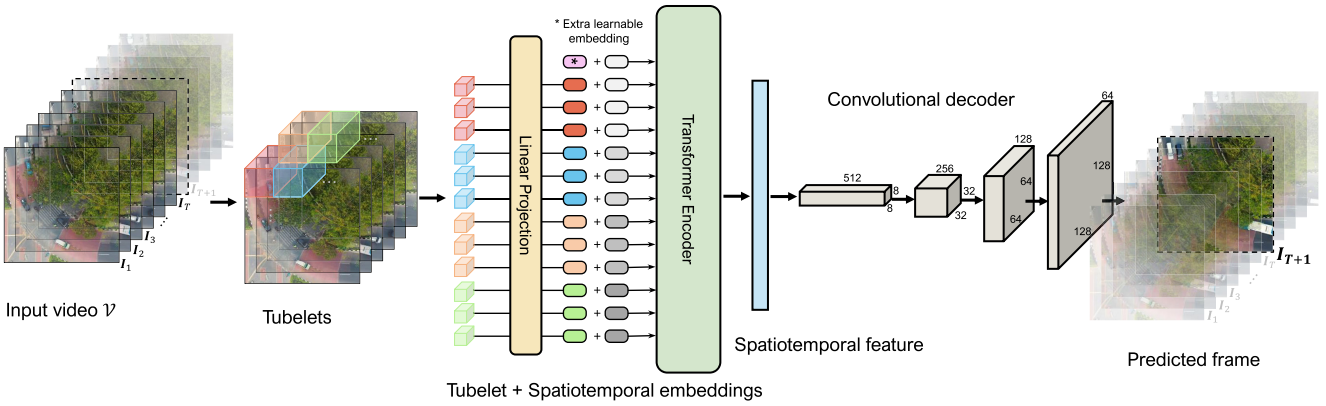


Fig. 3. Overview of ANDT. Our method treats a video as a sequence of tubelets and maps them into tubelet embeddings by linear projection. For preserving spatiotemporal information, the tubelet embeddings are added with learnable spatiotemporal position embeddings and then fed into a Transformer encoder to learn a spatiotemporal feature. The decoder is subsequently combined with the encoder for predicting the next frame based on the learned spatiotemporal representation.

In the test phase, the ℓ_2 distance between the predicted next frame $\hat{\boldsymbol{I}}_{T+1}$ and the true next frame $\boldsymbol{I}_{T+1}$ is calculated for identifying anomaly. The frames with relatively large ℓ_2 distances are deemed as anomalies.

B. Anomaly Detection With Transformers

We propose a method ANDT as the mapping $\mathcal{P}$. The Transformer [75] was originally proposed for sequence-to-sequence tasks in natural language processing (NLP), such as language translation. Its main idea is to use self-attention that enables the model to capture long-range dependencies in a whole sequence. We observe that a video is naturally a temporal sequence, but with spatial content. Therefore, we interpret a video as a sequence of tubelets and process them by a Transformer encoder to capture long-term spatiotemporal dependencies. Furthermore, a 3-D convolutional decoder is further attached for predicting the next frame based on the learned spatiotemporal relations. An overview of the model is depicted in Fig. 3.

Vision Transformer [76] performs tokenization by splitting an image into a sequence of small patches. In this work, since we deal with videos, we tokenize a video by extracting non-overlapping, spatiotemporal tubes. Specifically, the input video $\mathcal{V} \in \mathbb{R}^{T\times H\times W\times C}$ is split into a sequence of flattened 3-D tubelets $\boldsymbol{x}_k \in \mathbb{R}^{(n_t\cdot n_h\cdot n_w)\times(t\cdot h\cdot w\cdot C)}$, where (H, W) is the spatial size of video frames, C represents the number of channels, T denotes the number of frames, (t, h, w) is the dimension of each tubelet, $n_t = [(T/t)]$, $n_h = [(H/h)]$, and $n_w = [(W/w)]$. $N = n_t \cdot n_h \cdot n_w$ is the number of tokens. Then, we map the tubelets into a K-dimensional latent space by a trainable linear projection with weights $\boldsymbol{E} \in \mathbb{R}^{(t\cdot h\cdot w\cdot C)\times K}$. By doing so, the spatiotemporal information can be preserved during the tokenization.

We also prepend a learnable embedding $\boldsymbol{x}_{cls}$ to the sequence of tubelet embeddings. It also serves as the output feature $\boldsymbol{p}$ of the Transformer encoder. Furthermore, to inject original spatiotemporal position information into our model, we add learnable spatiotemporal position embeddings $\boldsymbol{E}_{\text{pos}} \in \mathbb{R}^{(N+1)\times K}$ to the tubelet embeddings. The equations are shown as follows:

$$z_0 = \left[\boldsymbol{x}_{\text{cls}}; \boldsymbol{x}_k^1\boldsymbol{E}; \boldsymbol{x}_k^2\boldsymbol{E}; \ldots; \boldsymbol{x}_k^N\boldsymbol{E}\right] + \boldsymbol{E}_{\text{pos}}. \tag{3}$$

z_0 is subsequently fed into Transformer encoder layers, each consisting of two sublayers. The first is a multihead self-attention (MSA) mechanism, and the second is a simple multilayer perceptron (MLP). Layer normalization (LN) is applied before every sublayer, and residual connections are used in every sublayer. The Transformer encoder takes these embeddings as input and learns a spatiotemporal feature $\boldsymbol{p}$ via

$$z_l' = \text{MSA}(\text{LN}(z_{l-1})) + z_{l-1} \tag{4}$$

$$z_l = \text{MLP}\big(\text{LN}\big(z_l'\big)\big) + z_l' \tag{5}$$

$$\boldsymbol{p} = \text{LN}\big(z_L^0\big) \tag{6}$$

where $l = 1, \ldots, L$.

We leverage a convolutional decoder to predict the next frame $\hat{\boldsymbol{I}}_{i+1}$ based on the learned spatiotemporal feature $\boldsymbol{p}$. First, we leverage two fully connected layers to increase the dimension of $\boldsymbol{p}$ and then reshape it into a 3-D tensor of $8 \times 8 \times 512$. This size is associated with the number of convolutional layers in the decoder. Considering both computational complexity and reconstruction accuracy, we use a decoder with five convolutional layers and upsampling layers. It progressively reconstructs the next frame with the size of $256 \times 256 \times 3$ from the encoded feature tensor of $8 \times 8 \times 512$. In particular, we leverage a progressive upsampling strategy that utilizes upsampling layers and convolution layers alternately. The upsampling rate is restricted to $2\times$. The batch normalization and ReLU are applied after each convolution layer. This strategy enables our decoder to learn spatial dependencies and upsample the learned features in a progressive manner, which leads to a better reconstruction of details and boundaries.

V. Experiments

In this section, we present our experimental results. In Section V-A, we introduce the datasets used in experiments. Evaluation metrics are introduced in Section V-B. Next, several ablation studies are conducted to investigate the

effectiveness of our method, and we report their results in Section V-D. Moreover, in Section V-E, we provide a benchmark on the Drone-Anomaly dataset for anomaly detection in aerial videos by extensively validating existing methods, and we compare our method with these baseline models. In Section V-F, we assess the performance of our method on AU-AIR-Anomaly dataset and compare our method with other competitors. Finally, we visualize the learned features of our method in Section V-G.

A. Dataset

To evaluate the performance of our method, we use not only our Drone-Anomaly dataset but also the AU-AIR-Anomaly dataset [39]. A statistic of the two datasets can be found in Table II.

B. Evaluation Metrics

The receiver operation characteristic (ROC) is a popular evaluation matrix in anomaly detection, and it is calculated by gradually changing the threshold. In addition, we also use area under curve (AUC) for the performance evaluation. We leverage a strategy to determine a threshold that is used to calculate recall, precision, $F1$ score, and overall accuracy (OA). Specifically, we feed the training set into the trained model to obtain reconstruction errors for all training samples. The threshold is determined as the sum of the mean value and the standard deviation value of the reconstruction errors. We note that AUC is the primary metric, as it can comprehensively evaluate the performance of a method.

C. Competitors

We compare our network with several state-of-the-art anomaly detection models.

1) *Convolutional Autoencoder (CAE) [67]:* The CAE aims to leverage the convolutional encoder to map the input frames into a latent space to learn features. A convolutional decoder is then employed to reconstruct a frame based on the learned features. Its reconstruction error is used for detecting anomalies.
2) *Convolutional Variational Autoencoder (CVAE) [69]:* The CVAE introduces a regularization into the representation space. It utilizes a prior distribution over the latent space to encode normal instances. This prevents the overfitting problem and enables the generation of meaningful frames for anomaly detection.
3) *Self-Adversarial Variational Autoencoder (adVAE) [70]:* The adVAE assumes that both anomalous and normal prior distributions are Gaussian. It utilizes a self-adversarial mechanism that adds discrimination training objectives to the encoder and decoder.
4) *GANomaly [71]:* GANomaly leverages a conditional generative adversarial network (GAN) to learn high-dimensional visual representations. It employs an encoder–decoder–encoder architecture in the generator network to enable the model to learn discriminative features of normality.
5) *Skip-GANomaly [72]:* Skip-GANomaly employs a convolutional encoder–decoder architecture with skip connections to thoroughly capture the multiscale distribution of normality.
6) *Memory-Augmented Autoencoder (MemAE) [73]:* The MemAE introduces a memory block between the encoder and the decoder. It records prototypical normal patterns optimally and efficiently by the proposed sparse addressing strategy.
7) *Memory-Guided Normality for Anomaly Detection (MNAD) [74]:* MNAD uses a memory module to record multiple prototypes that represent diverse representations of normalities for unsupervised anomaly detection.
8) *Multiresolution Knowledge Distillation for Anomaly Detection (MKD) [33]:* MKD proposes to distill the knowledge of a pretrained expert network into another more compact network to concentrate solely on discriminative features that are helpful in distinguishing normality and anomaly.
9) *Self-Supervised Predictive Convolutional Attentive Block (SSPCAB) [34]:* SSPCAB uses a convolutional layer with dilated filters, where the center area of the receptive field is masked. The block learns to reconstruct the masked area using contextual information. It can be incorporated into various existing models. In this article, we equip it on the MNAD [74] model, which is still denoted SSPCAB.

D. Ablation Studies

We present a series of ablations for evaluating the effectiveness of our model. All of them are conducted on the *highway* scene with the most number of training and test frames.

1) Model Design: In the course of experiments, we find that the design of the Transformer encoder matters. Hence, we want to investigate different configurations and figure out optimal settings. Concretely, the following hyperparameters are taken into account: patch size, number of Transformer layers, number of attention heads, and MLP size. From Table III(a), it can be observed that the model with a patch size of 16×16 achieves better comprehensive performance. The patch size is actually associated with the extent to which the model excavates inner information in patches and spatiotemporal relations among patches. In Table III(b) and (c), we focus on self-attention and find that using two Transformer layers and six attention heads exhibits superior performance. MSA enables the model to integrate multiple temporal information from different representation patches. Also, the small number of Transformer layers ensures a relatively small computational complexity. Finally, MLP size determines the size of the output spatiotemporal feature of the Transformer encoder. In Table III(d), we can see that an MLP with a size of 4096 brings good results to our model, which could be caused by the improved information capacity of the spatiotemporal feature.

2) Prediction Versus Reconstruction: In our network, future frame prediction is an important strategy to learn temporal dependencies for effectively detecting anomalies. To evaluate how it affects the performance, we compare

TABLE III

ABLATIONS ON THE ANDT DESIGN. WE SHOW AUC, $F1$ SCORE, AND OA OF SEVERAL TRANSFORMER DESIGNS WITH DIFFERENT CONFIGURATIONS. THE BEST ACCURACIES ARE SHOWN IN BOLD. (a) PATCH SIZE. THE MODEL WITH 16×16 EXHIBITS SUPERIOR PERFORMANCE AND EFFECTIVELY PRESERVES THE SPATIOTEMPORAL INFORMATION OF THE INPUT VIDEO. (b) NUMBER OF TRANSFORMER LAYERS. THE NETWORK WITH 2 LAYERS ACHIEVES A BETTER PERFORMANCE AND ALSO HAS RELATIVELY SMALL COMPUTATIONAL COMPLEXITY. (c) NUMBER OF ATTENTION HEADS. THE MODEL WITH 6 ATTENTION HEADS HAS A OUTSTANDING PERFORMANCE AND IS ABLE TO LEARN LONG-TERM TEMPORAL FEATURES. (d) MLP SIZE. THE MLP WITH THE SIZE OF 4096 ACHIEVES A BETTER PERFORMANCE. LARGER SIZE MLP IMPROVES THE INFORMATION CAPACITY OF SPATIOTEMPORAL FEATURES

	AUC	F1	OA
8×8	60.54	57.13	53.93
16×16	**64.32**	63.51	**61.56**
32×32	62.78	59.46	57.20
64×64	64.07	**65.78**	60.14

(a)

	AUC	F1	OA
1	62.41	58.32	56.06
2	**67.48**	**68.53**	**63.29**
4	64.32	63.51	61.56
6	63.71	60.46	59.53

(b)

	AUC	F1	OA
2	63.56	61.47	60.75
4	67.48	**68.53**	63.29
6	**68.12**	67.40	**64.18**
8	66.24	65.83	62.61

(c)

	AUC	F1	OA
768	65.18	65.86	63.79
1536	67.46	66.51	63.38
3072	68.12	67.40	**64.18**
4096	**68.65**	**67.68**	63.87

(d)

TABLE IV

PREDICTION VERSUS RECONSTRUCTION. WE SHOW NUMERICAL RESULTS OF THREE DIFFERENT ANOMALY DETECTION STRATEGIES. THE BEST RESULTS ARE SHOWN IN BOLD

Model	AUC	Recall	Precision	F1 score	OA	Δ_s
Reconstruction-1[1]	62.1	64.9	60.3	62.5	60.6	0.16
Reconstruction-6[2]	66.7	64.4	62.5	63.4	63.1	0.19
Prediction-1[3]	**68.7**	**68.4**	**66.9**	**67.7**	**65.9**	**0.25**

[1] Reconstruction-1 is the strategy of inputting 1 frame and reconstructing itself.
[2] Reconstruction-6 is the strategy of inputting 6 consecutive frames and reconstructing themselves.
[3] Prediction-1 is the strategy of inputting 6 consecutive frames and predicting the next frame.

TABLE V

NUMBER OF INPUT FRAMES. WE REPORT THE PERFORMANCE OF OUR MODEL WITH A VARIANT NUMBER OF INPUT FRAMES. THE BEST ACCURACIES ARE SHOWN IN BOLD

	AUC	Recall	Precision	F1 score	OA
2	63.7	68.0	58.9	63.1	62.5
4	67.4	64.5	69.7	67.0	65.1
6	**68.7**	68.4	66.9	**67.7**	**65.9**
8	67.1	63.2	**71.5**	67.0	65.4
10	65.8	64.9	65.4	65.2	63.0
12	64.0	**70.3**	60.4	65.0	62.2

our prediction-based framework with a commonly used reconstruction-based methodology [31], [66], [67], [68], [69], [70], [71], [72], [73], [74]. More specifically, with the same network architecture, we consider the following models: 1) inputting one frame, reconstructing itself; 2) inputting six consecutive frames, reconstructing themselves; and 3) inputting six consecutive frames, predicting the next frame (i.e., the proposed method). We first report the results of these models in the five evaluation metrics. Then, we calculate the difference between the average anomaly score of normal frames and that of abnormal frames, represented by Δ_s. The network with a relatively large Δ_s is more capable of distinguishing abnormal frames from normal frames. All results are shown in Table IV. It can be seen that the prediction-based framework can achieve better results in AUC, recall, $F1$ score, OA, and Δ_s.

3) Number of Input Frames: We further investigate how the number of input frames affects the performance of our method. We evaluate the performance of ANDT with a variant number of input frames. The results are reported in Table V. We can see that the method with six input frames exhibits superior comprehensive performance. The performance of our model gradually gets better, as the number of input frames goes from 2 to 6 and then degrades with more input frames. This observation demonstrates that a few frames are not enough for modeling temporal context, but too many input frames bring a deteriorated performance.

E. Results on the Drone-Anomaly Dataset

We evaluate various baseline models on all scenes in our Drone-Anomaly dataset with standard evaluation protocols and offer a benchmark. The results are reported in Tables VI and VII. Also, we compare the proposed model with other competitors.

1) Highway: This scene presents various kinds of anomalous events, e.g., a cow herd walking on the street, an accidental car collision, and a road section covered by sand and dust. These different anomalous events make this scenario very challenging. Comparing with other competitors, our method achieves the best results in AUC (68.7%) and recall (68.4%). The main competitor in this scene is MemAE that also exhibits very good results in some metrics. However, its accuracy in AUC is relatively a bit low. Our method demonstrates the capability of detecting different anomalous events and even presents better performance than memory-based methods, such as MemAE and MNAD, that are specially designed to deal with various anomalies.

2) Crossroads: This scene focuses on distinguishing various anomalous behaviors of vehicles and persons, such as persons crossing the road irregularly and vehicles moving backward. In this scene, capturing temporal dynamics of persons and vehicles on the road is critical for identifying their anomalous behaviors. From the reported results in Table VII, our method achieves the best results in AUC (65.2%), precision (66.3%), $F1$ score (64.6%), and OA (65.8%). This is mainly because the Transformer encoder of our approach is able to effectively model long-term temporal relations for distinguishing anomalous moving directions of persons or vehicles. We visualize the prediction of our method on a video clip of this scenario in Fig. 4 (the third row), in which an anomalous event is that a person crosses the road not following the rule. We can observe that the traffic is hindered by the person crossing the road irregularly. In this case, dynamically sensing traffic speed is

TABLE VI

COMPARING OUR APPROACH AGAINST OTHER METHODS. WE COMPARE OUR ANDT WITH OTHER COMPETITORS ON HIGHWAY, CROSSROADS, BIKE ROUNDABOUT, AND VEHICLE ROUNDABOUT SCENES. THE BEST ACCURACIES ARE SHOWN IN BOLD

Model	Highway					Crossroads					Bike roundabout					Vehicle roundabout				
	AUC	*Recall*	*Precision*	*F1 score*	OA	*AUC*	*Recall*	*Precision*	*F1 score*	OA	*AUC*	*Recall*	*Precision*	*F1 score*	OA	*AUC*	*Recall*	*Precision*	*F1 score*	OA
CAE [67]	58.3	60.4	58.8	59.6	57.1	57.7	60.7	61.3	61.0	60.3	59.4	57.7	59.0	58.3	58.8	60.9	58.9	56.5	57.7	58.4
CVAE [69]	61.7	64.1	63.4	63.7	61.0	62.4	61.5	61.8	61.7	59.7	76.5	68.8	73.4	71.0	68.7	57.6	58.4	57.6	58.0	56.4
adVAE [70]	61.1	59.7	60.3	60.0	59.1	56.1	56.9	54.8	55.8	56.5	72.8	71.8	75.9	73.8	69.4	55.1	54.4	52.9	53.6	54.1
GANomaly [71]	62.7	65.1	62.9	64.0	61.5	58.9	58.5	57.2	57.9	59.2	71.7	70.2	77.5	73.7	69.3	55.1	58.6	55.7	57.1	54.0
Skip-GAN [72]	64.8	63.7	66.7	65.2	64.6	59.3	60.3	60.6	60.4	62.1	77.7	73.5	74.3	73.9	67.7	58.5	59.3	62.8	61.0	57.1
MemAE [73]	67.2	67.3	**68.2**	67.7	66.1	64.1	**63.8**	63.3	63.6	64.5	79.5	74.8	73.4	74.1	75.2	**64.1**	**60.8**	59.0	59.9	59.2
MNAD [74]	66.9	65.9	66.5	66.2	65.7	56.6	57.2	59.4	58.3	55.2	77.4	72.4	75.2	73.8	69.8	61.9	57.9	61.6	59.7	59.4
MKD [33]	64.3	62.8	65.3	64.0	63.9	63.5	63.4	61.2	62.3	63.7	74.8	70.6	75.1	72.8	73.2	62.7	59.7	63.7	61.6	58.7
SSPCAB [34]	67.8	67.5	69.7	**68.6**	**66.3**	60.4	60.7	61.9	61.3	60.4	76.8	74.6	76.0	75.3	70.4	62.3	59.7	63.8	**61.7**	**60.4**
ANDT	**68.7**	**68.4**	66.9	67.7	65.9	**65.2**	63.1	**66.3**	**64.6**	**65.8**	**82.2**	**78.5**	**79.0**	**78.8**	**76.7**	61.3	57.8	**64.1**	60.8	58.0

TABLE VII

COMPARING OUR APPROACH AGAINST OTHER METHODS. WE COMPARE OUR ANDT WITH OTHER COMPETITORS ON RAILWAY INSPECTION, SOLAR PANEL INSPECTION, AND FARMLAND INSPECTION SCENES. THE BEST ACCURACIES ARE SHOWN IN BOLD

Model	Railway inspection					Solar panel inspection					Farmland inspection				
	AUC	*Recall*	*Precision*	*F1 score*	OA	*AUC*	*Recall*	*Precision*	*F1 score*	OA	*AUC*	*Recall*	*Precision*	*F1 score*	OA
CAE [67]	61.2	59.7	54.8	57.1	56.7	62.9	**62.7**	65.3	**64.0**	60.2	77.1	79.2	72.6	75.8	74.5
CVAE [69]	59.1	**62.8**	**64.7**	62.2	59.3	57.5	57.3	57.1	57.2	58.4	78.4	**80.7**	77.1	**78.9**	**75.7**
adVAE [70]	62.1	56.2	57.9	57.1	56.4	**66.1**	58.6	60.9	59.8	60.5	73.8	77.9	76.6	77.3	72.6
GANomaly [71]	61.7	55.7	56.2	56.0	53.8	64.6	59.1	63.7	61.3	57.3	77.1	74.0	73.2	73.6	75.5
Skip-GAN [72]	**65.8**	60.7	64.6	**62.6**	60.3	65.7	58.8	57.5	58.1	60.2	71.7	75.4	73.3	74.3	72.6
MemAE [73]	58.9	58.0	58.4	58.2	58.0	65.8	62.1	57.6	59.8	57.7	74.1	79.7	**77.7**	78.7	74.4
MNAD [74]	58.0	61.3	56.1	58.6	57.1	64.7	58.6	58.0	58.3	59.6	78.6	78.5	74.2	76.3	74.5
MKD [33]	62.4	59.7	60.3	60.0	**60.8**	63.5	57.6	54.7	56.1	56.5	75.2	76.8	72.4	74.5	72.8
SSPCAB [34]	59.1	62.0	58.7	60.3	59.4	65.0	59.2	60.9	60.0	58.7	79.0	78.4	75.8	77.9	75.1
ANDT	59.4	60.7	61.3	61.0	57.4	64.2	61.2	**66.0**	63.5	**60.8**	**79.5**	76.9	77.6	77.2	73.5

crucial for the successful detection of anomalous events. The numerical results demonstrate the effectiveness of our model. For evaluating the performance of detecting different kinds of anomalous events, we group anomalies into two categories: person-related anomaly and vehicle-related anomaly. The AUC results of each anomalous event are reported in Table VIII. Compared with other methods, our approach achieves the best AUC results in both two kinds of anomalies.

3) Bike Roundabout: Only one type of anomaly, i.e., moving vehicle on the bike roundabout, is presented in this scene. However, more than one abnormal event may be present in the test video sequence. This scenario can verify whether a method is able to continuously detect all anomalous events in a test sequence. Our method exhibits superior performance. We also observe that memory-based methods have poor performance. The reason for this may be that some feature representations of abnormal video frames misidentified as normality are memorized in the memory space, which deteriorates the performance of these models in recognizing subsequent anomalous frames.

TABLE VIII

AUC RESULTS OF DIFFERENT KINDS OF ANOMALIES IN CROSSROADS. WE OFFER AUC RESULTS OF TWO KINDS OF ANOMALIES IN CROSSROADS. THE BEST ACCURACIES ARE SHOWN IN BOLD

Model	Crossroads	
	person-related	*vehicle-related*
CAE [67]	61.8	55.0
CVAE [69]	59.9	64.1
adVAE [70]	57.2	55.4
GANomaly [71]	52.0	63.5
Skip-GAN [72]	56.4	61.2
MemAE [73]	64.7	63.7
MNAD [74]	57.3	56.1
MKD [33]	62.7	64.3
SSPCAB [34]	58.7	62.1
ANDT	**65.8**	**64.8**

4) Vehicle Roundabout: Various anomalous events, such as traffic congestion and people crossing the road irregularly, are present in this scene. Memory-based and GAN-based

Fig. 4. Visualization of anomaly detection results of our method and a main competitor. We show frame-level anomaly scores (orange curves indicate ANDT, and blue curves denote MemAE). Ten frames of each video are shown, and anomalous frames are marked with borders. Rectangles are ground-truth data. A demo video is available at https://youtu.be/ancczYryOBY.

methods, namely, Skip-GAN, MemAE, and MNAD, show superior performance in this scene. Our model suffers from insufficient training data and performs relatively poor.

5) Railway Inspection: This scene presents only one kind of anomaly, i.e., obstacles on the railway. Determining the existence of obstacles on the railway is vital in practical

TABLE IX

COMPARING OUR APPROACH AGAINST OTHER METHODS ON THE AU-AIR-ANOMALY DATASET. WE COMPARE OUR ANDT WITH OTHER COMPETITORS ON AU-AIR DATASET. THE BEST ACCURACIES ARE SHOWN IN BOLD

Model	AUC	Recall	Precision	F1 score	OA
CAE [67]	69.3	70.2	64.7	67.3	66.4
CVAE [69]	70.8	63.7	72.1	67.6	67.1
adVAE [70]	72.2	70.7	74.9	72.7	70.6
GANomaly [71]	70.4	73.6	61.8	67.2	72.8
Skip-GAN [72]	74.8	60.8	84.1	70.6	72.1
MemAE [73]	81.4	**87.6**	74.8	80.7	**82.4**
MNAD [74]	78.4	76.9	79.4	78.1	76.2
MKD [33]	76.8	83.7	79.6	81.6	79.5
SSPCAB [34]	79.6	77.4	80.4	78.9	78.3
ANDT	**86.7**	80.7	**84.9**	**82.7**	82.0

applications. From the results in Table VII, there is no dominant method. The reason might be the insufficient training data (only 400 frames are available for training) cannot ensure that these models learn strong feature representations of normality.

6) Solar Panel Inspection: Two anomalies, unknown objects/animals and panel defects, appear in this scene. Our model achieves the best accuracies in precision (66.0%) and OA (60.8%) and provides relatively satisfactory results in this scenario.

7) Farmland Inspection: One type of anomaly, i.e., unidentified vehicles, exists in this scene. Searching anomalous objects is the goal in this scene. From experimental results, our network achieves the best accuracies in AUC (79.5%) and exhibits superior performance in searching anomalous objects.

In summary, our model exhibits superior performance in multiple scenes, including *highway*, *crossroads*, *bike roundabout*, and *farmland inspection*, in which many anomalous events with temporal dynamics exist. Specifically, in the *highway* scene, our method presents a better performance of detecting different anomalies than memory-based methods, i.e., MemAE and MNAD, which are specially designed to deal with various anomalies. This is because the global temporal receptive field enables our model to learn discriminative temporal representations of normality, which is used to effectively detecting different anomalies.

F. Results on the AU-AIR-Anomaly Dataset

Furthermore, we use the AU-AIR-Anomaly dataset [39] to validate the performance of our approach and other methods. Due to the non-availability of public ground-truth labels for anomalies in the AU-AIR-Anomaly dataset, following [39], we label four anomalous events: a car on a bike road, a person on a road, a parked van in front of a building, and a bicycle on a road. We report numerical results in Table IX. As we can see, our model has a superb performance and achieves the best accuracies in AUC (86.7%), precision (84.9%), and $F1$ score (82.7%). The scene of this dataset is highly similar to crossroads in our Drone-Anomaly dataset. Our network still exhibits stable and superior performance, which demonstrates its good generalization ability across different datasets.

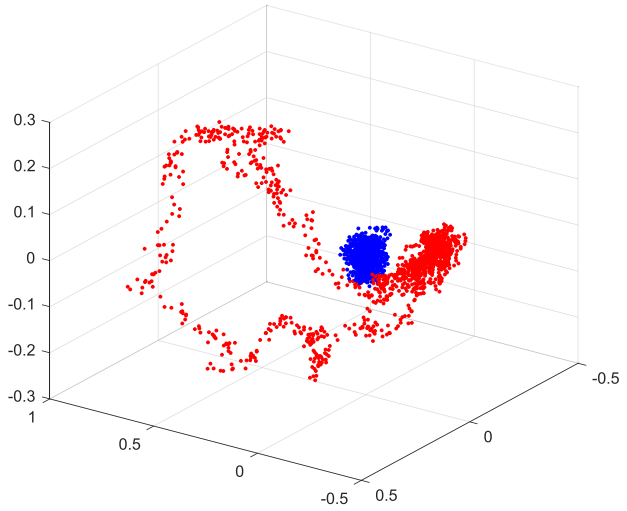


Fig. 5. Visualization of feature distribution. We visualize the distribution of the learned spatiotemporal features from the Transformer encoder on the *highway* scene. The features of normal frames are represented by blue points, and features of anomalous frames are red points.

Fig. 6. Sample frames in the scene of river inspection. We show four frames in this scene. All normal frames are marked with green borders.

G. Visualization of the Learned Features

We visualize, in Fig. 5, the distribution Transformer features of some randomly chosen test samples on the *crossroads* scene in the Drone-Anomaly dataset. We leverage a principal component analysis (PCA) to reduce the dimension of the features to 3. From Fig. 5, it can be seen that normal instances (blue points) are all concentrated in a relatively small area, while abnormal samples are far away from the blue cluster. This demonstrates that the spatiotemporal features learned by our model are very discriminative.

H. Discussion

To verify whether our method raises too many false alarms in practical applications that do not contain any anomalies, we collect a new scene, i.e., river inspection, which does not contain anomalous events. We use a DJI drone to inspect a normal river and collect an aerial video for this validation. We show four sample frames of the test data in Fig. 6. We report mean-squared reconstruction error (MSRE) values on training data and test data, and they are $\text{MSRE}_{\text{tra}} = 0.076$ and $\text{MSRE}_{\text{test}} = 0.078$. We can see that these two values are very close. Besides, we calculate false positive rate, $\text{FPR} = 0.0041$, which is very low. These mean that in scenes without any anomalies, our model also works well.

VI. CONCLUSION

In this article, we focus on detecting anomalous events in aerial videos. To this end, we create a new dataset, termed Drone-Anomaly, providing 37 training video sequences and

22 testing video sequences, covering seven real-world scenes, and providing various anomalous events. Based on this dataset, we offer a benchmark for this task. Moreover, we present a new baseline model, ANDT, which treats a video as a sequence of tubelets and leverages a Transformer encoder to learn a spatiotemporal feature. Afterward, a decoder is combined with the encoder for predicting the next frame based on the learned spatiotemporal representation. Also, we conduct extensive ablation studies for validating the effectiveness of our network. Moreover, we compare our model with other baselines. The experimental results demonstrate its outstanding performance. In the future, we will focus on spatiotemporally detecting anomalous events in aerial videos.

REFERENCES

[1] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection: A review," *ACM Comput. Surv.*, vol. 54, no. 2, pp. 1–38, Mar. 2021.

[2] X. Xie, Q. Lu, D. Rodenas-Herraiz, A. K. Parlikad, and J. M. Schooling, "Visualised inspection system for monitoring environmental anomalies during daily operation and maintenance," *Eng., Construct. Architectural Manag.*, vol. 27, no. 8, pp. 1835–1852, Jul. 2020.

[3] N. Neto and J. de Brito, "Validation of an inspection and diagnosis system for anomalies in natural stone cladding (NSC)," *Construct. Building Mater.*, vol. 30, pp. 224–236, May 2012.

[4] J. T. Zhou, J. Du, H. Zhu, X. Peng, Y. Liu, and R. S. M. Goh, "Anomalynet: An anomaly detection network for video surveillance," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 10, pp. 2537–2550, Oct. 2019.

[5] C. E. Au, S. Skaff, and J. J. Clark, "Anomaly detection for video surveillance applications," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, 2006, pp. 888–891.

[6] S. Tariq, H. Farooq, A. Jaleel, and S. M. Wasif, "Anomaly detection with particle filtering for online video surveillance," *IEEE Access*, vol. 9, pp. 19457–19468, 2021.

[7] B. S. Morse, D. Thornton, and M. A. Goodrich, "Color anomaly detection and suggestion for wilderness search and rescue," in *Proc. 7th Annu. ACM/IEEE Int. Conf. Hum.-Robot Interact. (HRI)*, Mar. 2012, pp. 455–462.

[8] M. T. Eismann, A. D. Stocker, and N. M. Nasrabadi, "Automated hyperspectral cueing for civilian search and rescue," *Proc. IEEE*, vol. 97, no. 6, pp. 1031–1055, Jun. 2009.

[9] S. Garg, K. Kaur, S. Batra, G. Kaddoum, N. Kumar, and A. Boukerche, "A multi-stage anomaly detection scheme for augmenting the security in IoT-enabled applications," *Future Gener. Comput. Syst.*, vol. 104, pp. 105–118, Mar. 2020.

[10] R. Yang, D. Qu, Y. Gao, Y. Qian, and Y. Tang, "NLSALog: An anomaly detection framework for log sequence in security management," *IEEE Access*, vol. 7, pp. 181152–181164, 2019.

[11] J. Yang, C. Zhou, S. Yang, H. Xu, and B. Hu, "Anomaly detection based on zone partition for security protection of industrial cyber-physical systems," *IEEE Trans. Ind. Electron.*, vol. 65, no. 5, pp. 4257–4267, May 2018.

[12] V. A. Desnitsky, I. V. Kotenko, and S. B. Nogin, "Detection of anomalies in data for monitoring of security components in the Internet of Things," in *Proc. 18th Int. Conf. Soft Comput. Meas. (SCM)*, May 2015, pp. 189–192.

[13] L. Mou, Y. Hua, P. Jin, and X. X. Zhu, "ERA: A data set and deep learning benchmark for event recognition in aerial videos [software and data sets]," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 4, pp. 125–133, Dec. 2020.

[14] M. Kiani, "Optimal image smoothing and its applications in anomaly detection in remote sensing," 2020, *arXiv:2003.08210*.

[15] M. A. Dias, E. A. D. Silva, S. C. D. Azevedo, W. Casaca, T. Statella, and R. G. Negri, "An incongruence-based anomaly detection strategy for analyzing water pollution in images from remote sensing," *Remote Sens.*, vol. 12, no. 1, p. 43, Dec. 2019.

[16] Q. Liu *et al.*, "Unsupervised detection of contextual anomaly in remotely sensed data," *Remote Sens. Environ.*, vol. 202, pp. 75–87, Dec. 2017.

[17] N. Wang, B. Li, Q. Xu, and Y. Wang, "Automatic ship detection in optical remote sensing images based on anomaly detection and SPP-PCANet," *Remote Sens.*, vol. 11, no. 1, p. 47, Dec. 2018.

[18] F. Yang, Q. Xu, B. Li, and Y. Ji, "Ship detection from thermal remote sensing imagery through region-based deep forest," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 3, pp. 449–453, Mar. 2018.

[19] D. W. J. Stein, S. G. Beaven, L. E. Hoff, E. M. Winter, A. P. Schaum, and A. D. Stocker, "Anomaly detection from hyperspectral imagery," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 58–69, Jan. 2002.

[20] C.-I. Chang and S.-S. Chiang, "Anomaly detection and classification for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 6, pp. 1314–1325, Jun. 2002.

[21] X. Kang, X. Zhang, S. Li, K. Li, J. Li, and J. A. Benediktsson, "Hyperspectral anomaly detection with attribute and edge-preserving filters," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5600–5611, Oct. 2017.

[22] Y. Yuan, D. Ma, and Q. Wang, "Hyperspectral anomaly detection by graph pixel selection," *IEEE Trans. Cybern.*, vol. 46, no. 2, pp. 3123–3134, Dec. 2015.

[23] Y. Liu, Q. Li, Y. Yuan, Q. Du, and Q. Wang, "ABNet: Adaptive balanced network for multiscale object detection in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2021.

[24] S. Wang, X. Wang, L. Zhang, and Y. Zhong, "Auto-AD: Autonomous hyperspectral anomaly detection network based on fully convolutional autoencoder," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[25] X. Zhang, X. Ma, N. Huyan, J. Gu, X. Tang, and L. Jiao, "Spectral-difference low-rank representation learning for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10364–10377, Dec. 2021.

[26] W. Xie, X. Zhang, Y. Li, J. Lei, J. Li, and Q. Du, "Weakly supervised low-rank representation for hyperspectral anomaly detection," *IEEE Trans. Cybern.*, vol. 51, no. 8, pp. 3889–3900, Aug. 2021.

[27] I. Golan and R. El-Yaniv, "Deep anomaly detection using geometric transformations," 2018, *arXiv:1805.10917*.

[28] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3379–3388.

[29] S. Zhai, Y. Cheng, W. Lu, and Z. Zhang, "Deep structured energy based models for anomaly detection," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016, pp. 1100–1109.

[30] A. Zimek, E. Schubert, and H.-P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data," *Stat. Anal. Data Mining*, vol. 5, no. 5, pp. 363–387, Oct. 2012.

[31] B. Zong *et al.*, "Deep autoencoding Gaussian mixture model for unsupervised anomaly detection," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–19.

[32] W. Lin, J. Gao, Q. Wang, and X. Li, "Learning to detect anomaly events in crowd scenes from synthetic data," *Neurocomputing*, vol. 436, pp. 248–259, May 2021.

[33] M. Salehi, N. Sadjadi, S. Baselizadeh, M. H. Rohban, and H. R. Rabiee, "Multiresolution knowledge distillation for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14902–14912.

[34] N.-C. Ristea *et al.*, "Self-supervised predictive convolutional attentive block for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13576–13586.

[35] D. Cavaliere, A. Saggese, S. Senatore, M. Vento, and V. Loia, "Empowering UAV scene perception by semantic spatio-temporal features," in *Proc. IEEE Int. Conf. Environ. Eng. (EE)*, Mar. 2018, pp. 1–6.

[36] D. Yang, K. Ozbay, K. Xie, H. Yang, F. Zuo, and D. Sha, "Proactive safety monitoring: A functional approach to detect safety-related anomalies using unmanned aerial vehicle video data," *Transp. Res. C, Emerg. Technol.*, vol. 127, Jun. 2021, Art. no. 103130.

[37] D. Cavaliere, V. Loia, A. Saggese, S. Senatore, and M. Vento, "A human-like description of scene events for a proper UAV-based video content analysis," *Knowl.-Based Syst.*, vol. 178, pp. 163–175, Aug. 2019.

[38] M. Bonetto, P. Korshunov, G. Ramponi, and T. Ebrahimi, "Privacy in mini-drone based video surveillance," in *Proc. IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Sep. 2015, pp. 1–6.

[39] I. Bozcan and E. Kayacan, "Context-dependent anomaly detection for low altitude traffic surveillance," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 224–230.

[40] I. S. Reed and X. Yu, "Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 38, no. 10, pp. 1760–1770, Oct. 1990.

[41] L. Wei and D. Qian, "Collaborative representation for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1463–1474, Mar. 2015.

[42] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014.

[43] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sensors*, vol. 2015, pp. 1–12, Jan. 2015.

[44] S. Matteoli, M. Diani, and G. Corsini, "A tutorial overview of anomaly detection in hyperspectral images," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 25, no. 7, pp. 5–28, Jul. 2010.

[45] A. Banerjee, P. Burlina, and C. Diehl, "A support vector method for anomaly detection in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 8, pp. 2282–2291, Aug. 2006.

[46] B. Du and L. Zhang, "Random-selection-based anomaly detector for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 5, pp. 1578–1589, May 2011.

[47] H. Kwon and N. M. Nasrabadi, "Kernel RX-algorithm: A nonlinear anomaly detector for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 2, pp. 388–397, Feb. 2005.

[48] A. Schaum, "Hyperspectral anomaly detection beyond RX," in *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XIII*, vol. 6565. Orlando, FL, USA: SPIE, 2007, pp. 502–656. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-SPIE/6565.toc

[49] D. R. Shaw and F. S. Kelley, "Evaluating remote sensing for determining and classifying soybean anomalies," *Precis. Agricult.*, vol. 6, no. 5, pp. 421–429, Oct. 2005.

[50] S. Liu *et al.*, "Detection of geothermal anomaly areas with spatio-temporal analysis using multitemporal remote sensing data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4866–4878, 2021.

[51] A. Chemura, O. Mutanga, and T. Dube, "Integrating age in the detection and mapping of incongruous patches in coffee (coffea arabica) plantations using multi-temporal landsat 8 NDVI anomalies," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 57, pp. 1–13, May 2017.

[52] R. Lasaponara, "On the use of principal component analysis (PCA) for evaluating interannual vegetation anomalies from SPOT/VEGETATION NDVI temporal series," *Ecol. Model.*, vol. 194, no. 4, pp. 429–434, Apr. 2006.

[53] A. Lanorte, T. Manzi, G. Nolè, and R. Lasaponara, "On the use of the principal component analysis (PCA) for evaluating vegetation anomalies from LANDSAT-TM NDVI temporal series in the basilicata region (Italy)," in *Proc. Int. Conf. Comput. Sci. Appl.*, 2015, pp. 204–216.

[54] P. Malhotra, L. Vig, G. M. Shroff, and P. Agarwal, "Long short term memory networks for anomaly detection in time series," in *Proc. Eur. Symp. Artif. Neural Netw., Comput. Intell. Mach. Learn. (ESANN)*, 2015, pp. 1–6.

[55] Y. He and J. Zhao, "Temporal convolutional networks for anomaly detection in time series," *J. Phys., Conf.*, vol. 1213, no. 4, pp. 042–050, 2019.

[56] Y. Zhang, Y. Chen, J. Wang, and Z. Pan, "Unsupervised deep anomaly detection for multi-sensor time-series signals," *IEEE Trans. Knowl. Data Eng.*, early access, Aug. 4, 2021, doi: 10.1109/TKDE.2021.3102110.

[57] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," in *Proc. Int. Symp. Neural Netw.*, 2017, pp. 189–196.

[58] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 733–742.

[59] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked RNN framework," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 341–349.

[60] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe, "Abnormal event detection in videos using generative adversarial nets," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 1–5.

[61] Y. Zhao, B. Deng, C. Shen, Y. Liu, H. Lu, and X.-S. Hua, "Spatio-temporal autoencoder for video anomaly detection," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1933–1941.

[62] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional LSTM for anomaly detection," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2017, pp. 439–444.

[63] J. Ryan Medel and A. Savakis, "Anomaly detection in video using predictive convolutional long short-term memory networks," 2016, *arXiv:1612.00390*.

[64] I. Bozcan and E. Kayacan, "AU-AIR: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 8504–8510.

[65] P. Jin, L. Mou, G.-S. Xia, and X. X. Zhu, "Anomaly detection in aerial videos via future frame prediction networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2021, pp. 8237–8240.

[66] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," *Special Lect. IE*, vol. 2, pp. 1–18, Dec. 2015.

[67] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *Proc. Int. Conf. Artif. Neural Netw. (ICANN)*, Jun. 2011, pp. 52–59.

[68] J. E. Fowler and Q. Du, "Anomaly detection and reconstruction from random projections," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 184–195, Jan. 2012.

[69] D. T. Nguyen, Z. Lou, M. Klar, and T. Brox, "Anomaly detection with multiple-hypotheses predictions," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 4800–4809.

[70] X. Wang, Y. Du, S. Lin, P. Cui, Y. Shen, and Y. Yang, "AdVAE: A self-adversarial variational autoencoder with Gaussian anomaly prior knowledge for anomaly detection," *Knowl.-Based Syst.*, vol. 190, Feb. 2020, Art. no. 105187.

[71] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "GANomaly: Semi-supervised anomaly detection via adversarial training," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2018, pp. 622–637.

[72] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Skip-GANomaly: Skip connected and adversarially trained encoder–decoder anomaly detection," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.

[73] D. Gong *et al.*, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1705–1714.

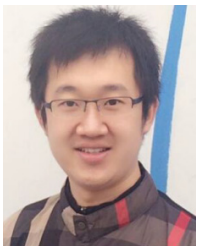
[74] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14372–14381.

[75] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 1–11.

[76] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020, pp. 1–22.

Pu Jin (Member, IEEE) received the bachelor's degree in electronic information science and technology from Wuhan University, Wuhan, China, in 2017, and the double master's degrees in Earth-oriented space science and technology (ESPACE) and photogrammetry and remote sensing from the Technical University of Munich (TUM), Munich, Germany, and Wuhan University, in 2020 and 2021, respectively. He is currently pursuing the Ph.D. degrees with the German Aerospace Center (DLR), Weßling, Germany, and TUM.

His research interests include remote sensing, computer vision, and deep learning, especially their applications in remote sensing.



Lichao Mou received the bachelor's degree in automation from the Xi'an University of Posts and Telecommunications, Xi'an, China, in 2012, the master's degree in signal and information processing from the University of Chinese Academy of Sciences (UCAS), Beijing, China, in 2015, and the Dr.Ing. degree from the Technical University of Munich (TUM), Munich, Germany, in 2020.

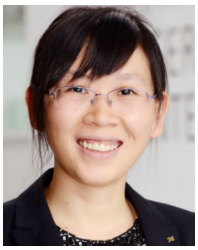
In 2015, he spent six months at the Computer Vision Group, University of Freiburg, Freiburg im Breisgau, Germany. In 2019, he was a Visiting Researcher with the Cambridge Image Analysis Group (CIA), University of Cambridge, Cambridge, U.K. He is currently a Guest Professor with the Munich Future Artificial Intelligence (AI) Lab—Artificial Intelligence for Earth Observation (AI4EO): Reasoning, Uncertainties, Ethics and Beyond, TUM, and the Head of the Visual Learning and Reasoning Team, Department of EO Data Science, Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Weßling, Germany. Since 2019, he has been a Research Scientist with the Institut für Methodik der Fernerkundung (IMF) im DLR, Weßling, and an AI Consultant for the Helmholtz Artificial Intelligence Cooperation Unit (HAICU), Munich.

Dr. Mou was a recipient of the First Place in the 2016 IEEE Geoscience and Remote Sensing Society (GRSS) Data Fusion Contest and finalists for the Best Student Paper Award at the 2017 Joint Urban Remote Sensing Event and the 2019 Joint Urban Remote Sensing Event.

Gui-Song Xia (Senior Member, IEEE) received the Ph.D. degree in image processing and computer vision from CNRS LTCI, Télécom ParisTech, Paris, France, in 2011.

From 2011 to 2012, he has been a Post-Doctoral Researcher with the Centre de Recherche en Mathématiques de la Decision, CNRS, Paris Dauphine University, Paris, for one and a half years. He was a Visiting Scholar with the Département de Mathématiques et Applications (DMA), École Normale Supérieure (ENS), Paris, for two months, in 2018. He is currently a Full Professor with Wuhan University, Wuhan, China, where he is leading a group involved in computer vision and photogrammetry. His research interests include mathematical modeling of images and videos, structure from motion, perceptual grouping, and remote sensing image understanding.

Dr. Xia serves on the editorial boards of several journals, including the *ISPRS Journal of Photogrammetry and Remote Sensing*, *Pattern Recognition*, *Signal Processing: Image Communications*, the *Journal of Remote Sensing*, and *Frontiers in Computer Science: Computer Vision*.



Xiao Xiang Zhu (Fellow, IEEE) received the M.Sc., Dr.-Ing., and Habilitation degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She was a Guest Scientist or a Visiting Professor with the Italian National Research Council (CNR-IREA), Naples, Italy, Fudan University, Shanghai, China, The University of Tokyo, Tokyo, Japan, and the University of California, Los Angeles, Los Angeles, CA, USA, in 2009, 2014, 2015, and 2016, respectively. She is currently the Professor with the Data Science in Earth Observation (former: Signal Processing in Earth Observation), TUM, and the Head of the Department of EO Data Science, Remote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany. Since 2019, she has been a Co-Coordinator of the Munich Data Science Research School (www.mu-ds.de), Munich, and has been the Head of the Helmholtz Artificial Intelligence—Research Field Aeronautics, Space and Transport, Munich. Since 2020, she has been the Director of the International Future Artificial Intelligence (AI) Lab—Artificial Intelligence for Earth Observation (AI4EO): Reasoning, Uncertainties, Ethics and Beyond, Munich. Since 2020, she has also been the Co-Director of the Munich Data Science Institute (MDSI), TUM. She is currently a Visiting AI Professor with the ESA's Phi-Lab, Frascati, Italy. Her research interests include remote sensing and Earth observation, signal processing, machine learning, and data science, with their applications in tackling societal grand challenges, e.g., global urbanization, Union Nations' Sustainable Development Goals (UN's SDGs), and climate change.

Dr. Zhu is a member of young academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities, the German National Academy of Sciences Leopoldina, and the Bavarian Academy of Sciences and Humanities. She serves in the scientific advisory board in several research organizations, among others the German Research Center for Geosciences (GFZ) and the Potsdam Institute for Climate Impact Research (PIK). She is an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and serves as an Area Editor responsible for the special issues of *IEEE Signal Processing Magazine*.