remote sensing

MDPI

Article

A Novel GAN-Based Anomaly Detection and Localization Method for Aerial Video Surveillance at Low Altitude

Danilo Avola [1,*], Irene Cannistraci [1], Marco Cascio [1], Luigi Cinque [1], Anxhelo Diko [1], Alessio Fagioli [1], Gian Luca Foresti [2], Romeo Lanzino [1], Maurizio Mancini [1], Alessio Mecca [2] and Daniele Pannone [1]

[1] Department of Computer Science, Sapienza University, 00198 Rome, Italy
[2] Department of Mathematics, Computer Science and Physics, University of Udine, 33100 Udine, Italy
* Correspondence: avola@di.uniroma1.it

Abstract: The last two decades have seen an incessant growth in the use of Unmanned Aerial Vehicles (UAVs) equipped with HD cameras for developing aerial vision-based systems to support civilian and military tasks, including land monitoring, change detection, and object classification. To perform most of these tasks, the artificial intelligence algorithms usually need to know, a priori, what to look for, identify. or recognize. Actually, in most operational scenarios, such as war zones or post-disaster situations, areas and objects of interest are not decidable a priori since their shape and visual features may have been altered by events or even intentionally disguised (e.g., improvised explosive devices (IEDs)). For these reasons, in recent years, more and more research groups are investigating the design of original anomaly detection methods, which, in short, are focused on detecting samples that differ from the others in terms of visual appearance and occurrences with respect to a given environment. In this paper, we present a novel two-branch Generative Adversarial Network (GAN)-based method for low-altitude RGB aerial video surveillance to detect and localize anomalies. We have chosen to focus on the low-altitude sequences as we are interested in complex operational scenarios where even a small object or device can represent a reason for danger or attention. The proposed model was tested on the UAV Mosaicking and Change Detection (UMCD) dataset, a one-of-a-kind collection of challenging videos whose sequences were acquired between 6 and 15 m above sea level on three types of ground (i.e., urban, dirt, and countryside). Results demonstrated the effectiveness of the model in terms of Area Under the Receiving Operating Curve (AUROC) and Structural Similarity Index (SSIM), achieving an average of 97.2% and 95.7%, respectively, thus suggesting that the system can be deployed in real-world applications.

Keywords: UAVs; computer vision; anomaly detection; aerial images; low-altitude; UMCD dataset; deep learning; GANs; AUROC; SSIM

Citation: Avola, D.; Cannistraci, I.; Cascio, M.; Cinque, L.; Diko, A.; Fagioli, A.; Foresti, G.L.; Lanzino, R.; Mancini, M.; Mecca, A.; et al. A Novel GAN-Based Anomaly Detection and Localization Method for Aerial Video Surveillance at Low Altitude. *Remote Sens.* **2022**, *14*, 4110. https://doi.org/10.3390/rs14164110

Academic Editor: Junjun Jiang

Received: 7 July 2022
Accepted: 19 August 2022
Published: 22 August 2022

1. Introduction

In recent years, computer vision techniques have become continuously more important to support a wide range of application areas, including foreground detection and/or background modeling for anti-intrusion and monitoring systems [1–5], object detection and target recognition for security and observation systems [6–10], human action and behavior recognition [11–16], assessment of progress of motor impairments by vision-based rehabilitation systems able to analyze body movements over time [17–19], or to support human-centered interfaces to drive advanced devices [20–22]. One of the application areas where computer vision has grown the most is without doubt the development of vision systems based on UAVs. In fact, as reported in the literature of the last 20 years, more and more UAVs equipped mostly with HD RGB cameras are commonly used to support innovative solutions in everyday life, such as precision agriculture [23,24], search and rescue [25,26], fire detection [27,28], and many others. Among these solutions, those relating to the aerial video surveillance of people, vehicles, and, events that occur on

the ground can be considered the most active research areas. In fact, on the one hand, an increasingly growing number of civilian applications need to perform monitoring missions able to interpret always more complex events, e.g., identification of flooded areas, detection of missing persons, and monitoring of quarantine zones; on the other hand, the needs of military applications are evolving quickly, requiring smart UAVs able to support a broad variety of dangerous missions, e.g., patrols, investigations, and protection of critical structures.

According to an in-depth analysis of the state of the art, traditional aerial video surveillance systems leverage pre-trained networks capable of recognizing and classifying a limited number of objects or events, which is not helpful under challenging circumstances in which unknown instances occur. Indeed, in [29], the authors present a robust multi-class object detection method combining a deep residual network, a feature pyramid structure, and a rotation region proposal network for extracting multiscale features and generating candidate-oriented bounding boxes surrounding only known targets in the aerial images. Again, in [30], the authors propose a novel Global Density Fused Convolutional Network (GDF-Net) composed of a backbone network, a global density model, and an object detection network, detecting and localizing only well-known target categories. Finally, in [31], the authors propose a pre-processing step and a pre-trained Convolutional Neural Network (CNN) to design a sea search and rescue system detecting already known objects of interest. Despite the excellent results obtained by these systems and their undoubted usefulness in critical situations, the recent trends focus on developing algorithms based on the Anomaly Detection (AD) concept. The latter moves its center of attention from detecting a set of classes known a priori to also detecting an eventually unknown set of objects or events that broadly differ from the others comprising the examined environment in terms of visual appearance and occurrences. In recent years, despite the increasing interest by the computer vision scientific community in AD applications, its usage for solving critical tasks (e.g., IED detection) in aerial surveillance from low-altitude video streams requires further investigations to consider all the possible challenges introduced by different types of ground (e.g., background clutter), variable size of objects (e.g., small objects with few pixels), and many others. Addressing such issues is the primary purpose of the present paper.

Starting from some of our previous experiences in designing UAV-based vision systems for aerial video surveillance [32–37] and from some interesting works in the use of the AD techniques in aerial images [38–41], in this paper, a novel two-branch GAN-based approach for low-altitude RGB aerial video streams to detect and localize anomalies in different types of ground for surveillance missions is presented. Notice that in this work, we are interested in detecting anomalies in particular operational scenarios such as analysis of the ground where military convoys travel to determine the presence of small dangerous objects (i.e., IEDs), analysis of the ground surrounding field hospitals to determine the presence of disguised small objects (i.e., traps), detection of small objects and artifacts that should not be in a specific place and, for this reason, are potentially dangerous, and many others. To test the proposed model, extensive experiments were conducted on the UMCD dataset [42]. To the best of our knowledge, this collection of video streams is the only one suitable to prove the effectiveness of the proposed model due to its distinctive characteristics. More specifically, the UMCD dataset is a collection of 50 challenging video sequences acquired at very low altitudes (i.e., between 6 and 15 m). Moreover, the videos are acquired with different parameters (e.g., speed, height) on different types of ground (i.e., urban, dirt, and countryside), introducing background clutter to stress the model localization capability. On the ground, there are natural and artificial structures (e.g., trees, buildings) to stress the robustness of the model as well as objects of common size (e.g., vehicles, persons) and very small objects (e.g., gas bottles, boxes, suitcases) to stress the detection ability of the model. Another unique characteristic of this dataset is the acquisition modality that acquires the videos on several of the same paths twice, with and without anomalous objects; the latter step is fundamental for the training stage in this kind of system. The metrics taken into

account to evaluate the proposed method were the AUROC and SSIM; the first was used to test the detection branch related to the binary classification (97.2%), whereas the second was used to evaluate the localization branch related to the measure of the similarity between the original image and the other generated ones using the GAN network (95.7%). In both cases, the outstanding obtained results have confirmed that the proposed system can be used in real-world applications. The main contributions of the paper can be summarized as follows:

- Designing a network architecture based on two GANs organized on parallel branches and intended for the modeling and learning of robust normal class data manifolds, which are crucial for increasing the performance and precision of detection and localization of eventually anomalous conditions;
- Detecting and localizing any anomalous element of interest in aerial videos at very low altitude (from 6 to 15 m), spanning from common items, e.g., cars or people, to undefined and challenging objects, e.g., IEDs, independently from their properties such as color, size, position, or shape, including elements never seen before;
- Presenting quantitative and qualitative experiments for the anomaly detection and localization tasks on the UMCD dataset, reaching outstanding results.

The paper is structured as follows. In Section 2, an overview of key works in anomaly detection and localization is reported. In Section 3, the novel two-branch GAN-based method for low-altitude RGB aerial video surveillance to detect and localize anomalies is detailed. In Section 4, quantitative and qualitative results obtained on the UMCD dataset are discussed. Finally, Section 5 contains the conclusions and proposes ideas for possible future improvements.

2. Related Work

AD popularity increased in recent years by proposing new and different strategies to improve the accuracy of automated systems. In general, an *anomaly* can be defined as an item, such as a person, an object, an animal, and so on, that is not expected to be in a specific environment. For example, the presence of a group of sheep in a farm area can be considered as a *normal* event, while, in the same context, an airplane will probably represent an anomaly. However, the same airplane, located in a hangar or the sky, can be considered *normal*. Given that, AD strategies aim at localizing and identifying all the items (e.g., persons, objects, animals) that are not expected to be found in a specific context. More specifically, AD aims to monitor an area of interest only by knowing what is considered *normal* for the application context. Automated AD systems usually provide significant benefits in several real-life scenarios. In security, for example, they can be used to monitor restricted/dangerous areas from intruders. In a military context, instead, they can help in retrieving unexploded bombs by localizing them from the distance, e.g., with drones. An overview of the more well-known AD approaches can be found in [43], while more recent surveys in [44–46] present novel proposals in this field. The recent widespread use of drones, also thanks to their cost reduction and technical improvements (e.g., flight time, automatic control, remote transmission), allows them to cover a significant part of AD applications, especially in large outdoor scenarios. Our proposal is focused on this specific context and exploits data captured by UAVs to detect possible anomalies at low-altitudes. For this reason, this section describes some recent state-of-the-art approaches for UAV-based AD, taking into account different operative contexts in order to provide an overview of interesting strategies in the various application areas.

In [47], an end-to-end method based on a deep one-class classification exploiting unsupervised generative learning is described. In this work, an event is made explicit by an optical flow and the original images coming from the UAV. The proposed strategy is focused on two different tasks. The first aims at maintaining the descriptive compactness of the normal event features. The last, instead, generates new optical flows directly from the data acquired by the UAV during the testing phase. This process speeds up the detection of possible anomalies and allows the system to fulfill the real-time constraint. To this

aim, their proposed network is an optical flow generator based on a deep CNN. More in detail, such a network does not compute the optical flows in a classical way. Instead, it fastens the process by exploiting a convolution/deconvolution-based neural network. This network is also able to extract compact features from both original images and optical flows. The authors also introduce a custom loss function for the training. It consists of the sum of three different loss functions, namely reconstruction loss, generation loss, and compactness loss. In this way, it is possible to achieve a more efficient classification of events. Their proposed strategy has been benchmarked on an in-house dataset composed of 1000 samples and on two publicly available datasets [48,49], providing remarkable results.

Another work focused on the use of AD techniques by UAVs is reported in [50], where the authors base their strategy on the use of four different sets of features. The first set relies on the deep features retrieved by GoogLeNet [51]; the second contains the local shape information extracted by the Histogram of Oriented Gradients (HOG) [52] from each region of the frames; the third is computed by exploiting the Principal Component Analysis (PCA) [53] on the previously mentioned features extracted by the HOG. Finally, the last is made up by the spatio-temporal features retrieved by the HOG3D [54] algorithm. Each feature set is separately fed to a One-Class Support Vector Machine (OC-SVM) classifier [55]. In this way, four different classifiers are trained. This work raises some interesting considerations. The first is that the second and the third sets of features produce a significantly worst classifier compared to the other two; in addition, the use of the PCA only improves the system's speed but not the performance. Finally, the best set seems to be the first one in which the GoogLeNet features achieved higher accuracy.

In [56], a strategy for AD based on future frame prediction with an encoder–decoder network is presented. The network proposed by the authors is exploited to extract spatio-temporal feature representations from the video frames. In their network, the last batch normalization and ReLU layers in the encoder part are removed to retain different representations, substituting L2 normalization layers in their place. Their proposed model is trained to predict a future frame receiving consecutive frames as input. Such a frame is then compared with the ground truth to detect an anomaly. The work proposed by the authors also exploits variational auto-encoders made up of an encoder and a decoder, thus optimizing the encoding/decoding scheme. The work reported in [39] instead proposes an unsupervised AD Deep Neural Network (DNN) for Aerial Surveillance. The proposed DNN aims at learning the distribution of the objects for each different environment according to GPS labels; at the same time, the network is constrained to have a continuous latent space to also learn the data distribution. The pipeline starts with a primary phase in which an off-the-shelf object detector (a MobileNetV2-SSDLite [57]) provides object annotations. The detector is exploited to compute the grid representations, which contain values according to a particular object's presence or absence. The network is fed with these representations jointly with the GPS coordinates. Experimental results demonstrate the effective use of the GPS coordinates, which significantly increase the precision of the network.

Moving to the AD by UAVs at low altitude, to the best of our knowledge, there is only one work addressing the topic [32], and it is not based on a deep learning strategy. It is worth noting that a change in the altitude acquisition may present several challenges due to the different sizes of the anomalies to detect. In addition, the altitude can also impact the visual features extracted from the scene. The just reported proposal is based on the extraction of a modified set of Haralick-based features [58,59] and exploits them to train an OC-SVM classifier. In the first step, the authors convert the input image in gray-scale and then split it into homogeneous patches; in the second step, they extract the Haralick-based features. In their work, the authors propose a spatial relation based on discretized circumferences of different ranges instead of the classical orientations (i.e., 0°, 45°, and so on). In the final step, the authors exploit the extracted features of each patch to train an OC-SVM classifier. Their proposal has been tested on the same dataset used in this work, i.e., the UMCD dataset [42], which contains low-altitude UAV-captured videos with and without different anomalies.

3. Methodology

The two-branch GAN-based neural network architecture, shown in Figure 1, was designed to detect and localize anomalies from RGB video streams acquired by UAVs at low altitude. Each model branch expands the GAN network [60,61] by leveraging the CNN architecture to handle visual data [62,63] and, at the same time, modeling and learning two different low-dimensional manifolds of video frames depicting anomaly-free scenarios; later, they are employed in detecting and localizing eventually abnormal conditions. Specifically, given RGB video frames, the top branch is a GAN serving as the detector and predicting whether or not they depict a normal scene, i.e., the presence of anomalous states. Instead, the bottom branch is a GAN operating as the localizer, producing an attention map, thus highlighting the abnormal elements within the frames when detected, i.e., localizing the anomalous regions in the scene. To this end, the network training follows a semi-supervised fashion, where the network inputs are the frames coming only from anomaly-free scenarios. In such a manner, with an anomalous frame as input, each element deviation and differentiation from the learned manifold can be recognized independently from the position, size, color, and shape during the testing phase. For illustration, the training and testing workflows are depicted in Figure 2a,b, respectively. Below, the proposed method and the related strategy are described in detail.

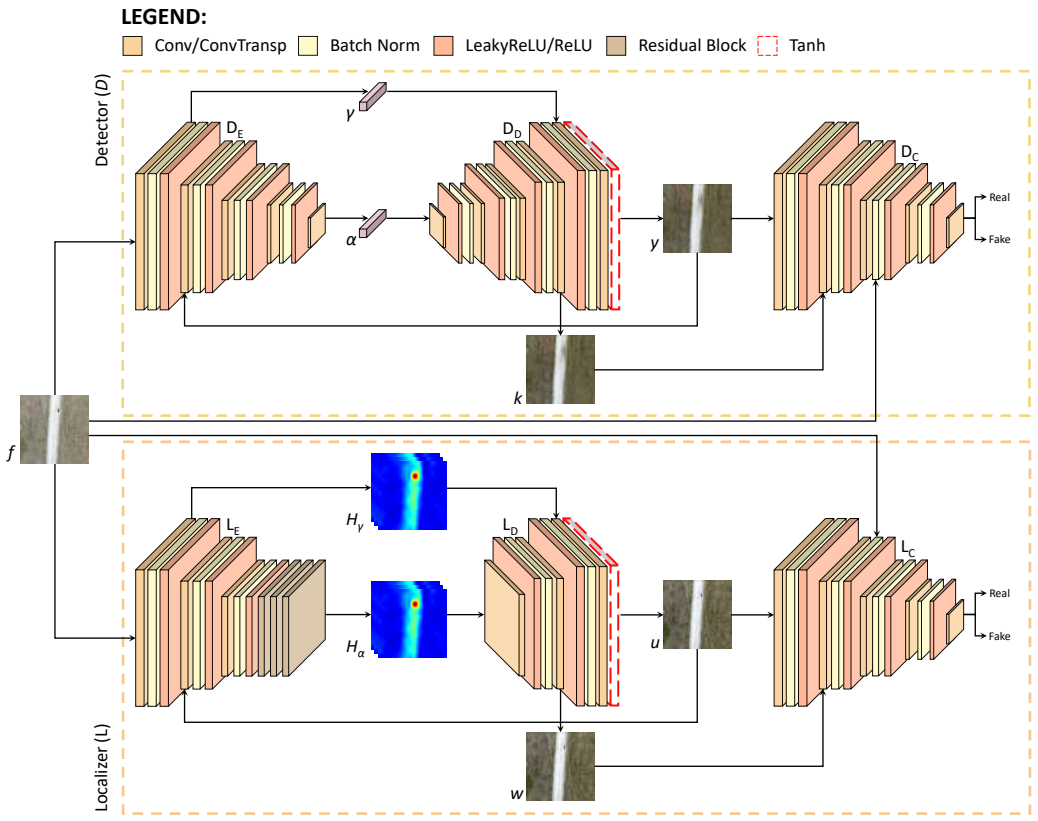


Figure 1. The proposed two-branch network architecture. Given an anomalous scene as input, the top GAN (i.e., the detector) detects if anomalies are present or not. The bottom GAN (i.e., the localizer), instead, localizes and highlights the anomalous regions when detected.

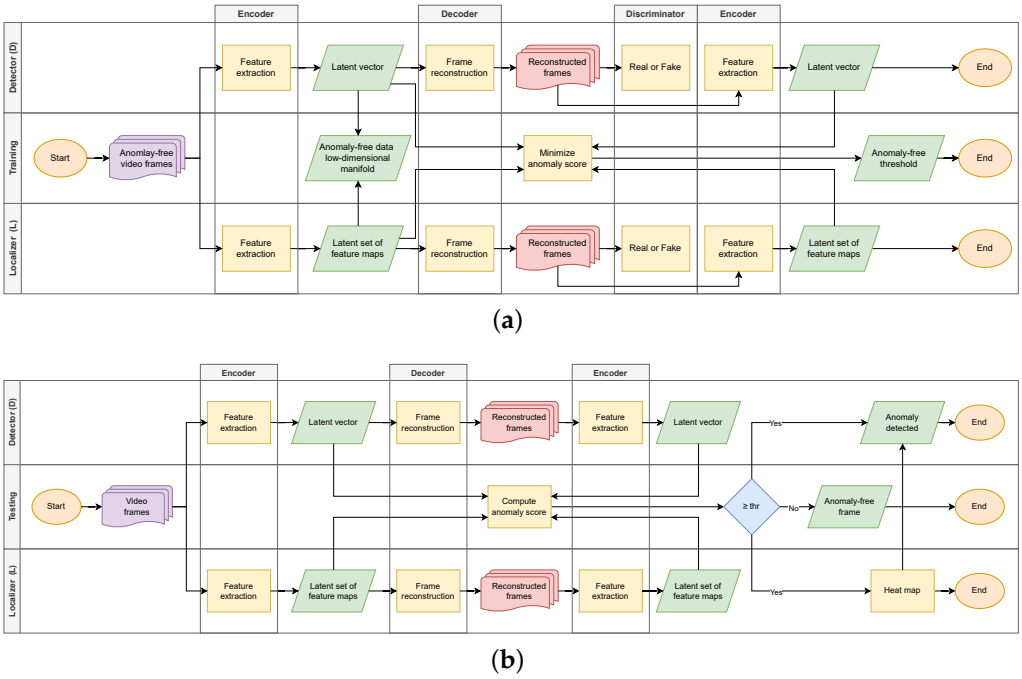


Figure 2. In (**a**,**b**), the detailed workflows required for detecting and localizing anomalies in RGB videos during the training and testing phases, respectively.

3.1. Anomaly Detection

In recent years, the GAN model has gained momentum for detecting anomalies over a set of data samples, including vision-based information [64,65], achieving promising results. This network architecture can approximate a target data distribution by generating artificial samples as if they were drawn from the target itself, i.e., learning the manifold of the observed data. Notice that this last property can be helpful under an image reconstruction setting to implement the detection of anomalies by exploiting the learned manifold of the normal class. Indeed, the just introduced property is employed by the top branch of the proposed network architecture, i.e., the detector D, to predict whether or not there are anomalies in the observed scene. In detail, the detector is a CNN-based GAN consisting of video frames encoder D_E, decoder D_D, and discriminator D_C components, where the encoder generates the latent vector for each frame, while the decoder reconstructs these frames from such low-dimensional representations produced by the encoder. The discriminator, instead, discriminates against all real and fake video frames. Specifically, both D_E and D_C models have similar structures comprising L convolutional layers with strided convolutions, each of them exploiting the batch normalization [66] to stabilize the training phase and the leaky rectified linear unit (leakyReLU) [67] activation function. However, the discriminator being a binary classifier, the sigmoid function is applied to the output of its last convolutional operation. The D_D model requires a reverse structure with respect to the D_E component. Therefore, it includes transposed convolutions and a rectified linear unit (ReLU) activation function replacing the convolutional operations and the leakyReLU activation, respectively, and the use of a hyperbolic tangent function in the last layer. More precisely, given a video sequence F, for each original frame $f \in F$, the encoder D_E produces the corresponding latent vector α. Afterward, the decoder D_D uses this low-dimensional representation to reconstruct the real frame as the fake image y, which is given as input to D_E,producing its corresponding latent vector γ with the same size of α. Finally, similarly to α, the vector γ is decoded by D_D to reconstruct another fake frame k, always reproducing the original frame f. Following this new training strategy, the detector can learn a more robust low-dimensional manifold of normal frames; indeed, when D_C classifies both y and k as real, it implies that the low-dimensional representation α is very informative to the point of fooling the discriminator not only with the reconstructed

frame y of the original sequence but even with the reconstruction of the fake frame in k. Formally, to accomplish the reconstruction goal, the low-dimensional manifold is learned applying the GAN adversarial function based on a zero-sum game. The latter is derived from the cross-entropy between original and reconstructed frames, as follows:

$$\begin{aligned} \mathcal{L}_f &= E_f[logD_C(f)], \\ \mathcal{L}_\alpha &= E_\alpha[log(1 - D_C(D_D(\alpha))], \\ \mathcal{L}_\gamma &= E_\gamma[log(1 - D_C(D_D(\gamma))], \\ \mathcal{L}_{adv} &= \min_{D_E, D_D} \max_{D_C} (\mathcal{L}_f + \mathcal{L}_\alpha + \mathcal{L}_\gamma), \end{aligned} \tag{1}$$

where $D_D(\alpha)$ and $D_D(\gamma)$ are the reconstructed frames (i.e., y and k) from real and fake data, respectively. Moreover, $D_C(f)$, $D_C(D_D(\alpha))$, and $D_C(D_D(\gamma))$ indicate the discriminator-estimated probability of a given frame being real. Finally, E_f, E_α, and E_γ are the expected values over all the real and fake frames.

During the training phase, the detector only learns how to reconstruct frames depicting the normal scene from their latent vectors, i.e., the manifold of normal data. To this end, since the fake frames must reproduce the original anomaly-free video sequence, the mean squared error (MSE) is defined between the real f and reconstructed y frames, with size (m, n), as follows:

$$MSE_r^D = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} [f(i,j) - y(i,j)]^2, \tag{2}$$

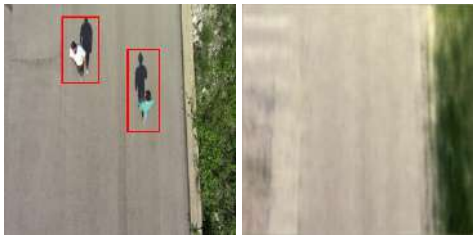
where $f(i,j)$ and $y(i,j)$ indicate a pixel in the original and reconstructed images, respectively. To improve the input frame low-dimensional representation quality, it is ensured that the original video can also be reproduced starting from fake frames by defining a similar constraint among the vectors α and γ as follows:

$$MSE_e^D = \frac{1}{|\alpha|} \sum_{i=1}^{|\alpha|} (\alpha_i - \gamma_i)^2, \tag{3}$$

where $|\alpha|$ is the latent vector size, which is the same for both representations. Finally, the training objective of the detector D can be computed via the following weighted function:

$$\mathcal{L}_D = w_r^D \, MSE_r^D + w_e^D \, MSE_e^D, \tag{4}$$

where w_r^D and w_e^D, with $w_r^D > w_e^D$, are weighting parameters adjusting the impact of individual losses on the overall loss function. By binding the input and reconstructed frames, as well as their corresponding low-dimensional vectors via Equation (3), the detector D naturally tries to fix any anomaly during the input frame reconstruction of an anomalous scene at the expense of the reconstruction quality itself, as illustrated in Figure 3. Following this rationale, the Euclidean distance among the low-dimensional representations of an anomaly input frame and its repaired reconstruction can be used to detect anomalies, as they result in highly different values. Therefore, the distance between latent vectors can be considered as an anomaly score, indicating the presence of anomalies when it is equal to or greater than a specific threshold, automatically learned when training the model by applying the Youden's J statistic [68].



(**a**) (**b**)

Figure 3. Example of image reconstruction of a frame depicting abnormal elements during the test phase. In (**a**) the abnormal frame; in (**b**) the reconstructed image of the repaired frame (i.e., removal of people in the middle of the street). Due to the anomaly correction, (**b**) has a low quality, resulting in a large Euclidean distance between (**a**,**b**) latent representations that indicates the detection of an anomaly.

3.2. Anomaly Localization

The great capability of GAN to approximate data distributions among the others can be employed for localizing anomalous elements in visual-based anomaly detection applications [69,70]. To this end, the network ability can be refined to learn features by modeling more abstract information from training samples, achieving pixel-wise localization. Therefore, the strategy is to understand the details appearing frequently in normal frames, employing them to predict abnormal areas within the scene. Moreover in this case, this is achieved by learning the manifold of normal data while preserving spatial information. Indeed, the bottom branch of the proposed network architecture, i.e., the localizer L, has similar components and structures of the model D described in Section 3.1. However, for the localizer, the low-dimensional representation is a set of 2D feature maps that can be used to define an heat map localizing regions of interest in the image. To achieve this, different from D_E, the last convolution in the localizer encoder L_E is replaced with three residual blocks [71] to improve localization performance [72] having more abstract representations of input frames. Precisely, similar to the detector branch, the L model learns to reconstruct only the frames depicting a normal scene but exploiting the instructive latent set of feature maps rather than a single vector, i.e., the learned manifold of normal data containing spatial information.

Formally, given a video sequence F, each original frame $f \in F$ is used by the encoder component L_E to generate a latent set H_α of 2D feature maps. Afterwards, the decoder component L_D uses all these low-dimensional feature maps to reproduce the real frame as the fake image u, which is given as input to L_E to produce another latent set H_γ of feature maps with the same shape and size of H_α. Finally, the latter is utilized by L_D to produce another fake frame w reproducing the original frame f. The localizer discriminator L_C, similar to D_C, is tasked with discriminating all real and fake frames. When L_C classifies both frames u and w as real, as discussed for the detector branch and the new training strategy, the localizer L learns a very informative latent set of feature maps H_α and a more robust low-dimensional manifold of normal frames. To reproduce frames depicting anomaly-free scenarios, even if the latent space is different from the detector, a similar constraint is required between real and fake frames, as follows:

$$MSE_r^L = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} [f(i,j) - u(i,j)]^2, \tag{5}$$

where $f(i,j)$ and $u(i,j)$ indicate a pixel in the original and reconstructed images with size (m,n), respectively. Even in this case, to improve the low-dimensional feature maps quality to the point that the original video sequence can be reproduced also starting from fake frames, the H_α and H_γ representations are bound via:

$$MSE_e^L = \frac{1}{N}\frac{1}{2s}\sum_{n=1}^{N}\sum_{i=1}^{s}\sum_{j=1}^{s}[H_\alpha^n(i,j) - H_\gamma^n(i,j)]^2, \tag{6}$$

where N is the total number of features maps within each se, s is the n-th feature map size, and $H_\alpha^n(i,j)$ and $H_\gamma^n(i,j)$ are the elements in position (i,j) for the feature map $n \in N$ related to the original and reconstructed frames, respectively. Therefore, the training objective for the localizer branch L is the following weighted loss function:

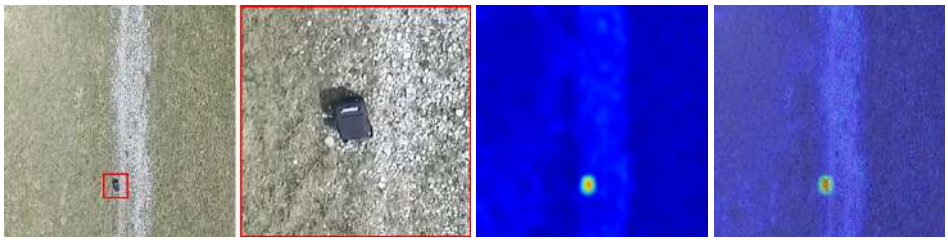
$$\mathcal{L}_L = w_r^L\ MSE_r^L + w_e^L\ MSE_e^L, \tag{7}$$

where w_r^L and w_e^L, with $w_r^L > w_e^L$, are weighting parameters adjusting the impact of individual losses on the overall loss function. Forcing these constraints, the L model becomes capable of capturing spatial differences, i.e., anomalous elements, within the latent set of feature maps of abnormal scenarios since the set H from anomalous videos will differ from reconstructed ones. Following this strategy, when an anomaly scene is given as input, abnormal regions are identified by producing a heat map $\mathcal{M}$ using the sets of low-dimensional feature maps obtained from the original and reconstructed frames. This heat map highlights the most significant areas of videos to observe for noticing the anomalies. It is computed by averaging and min–max normalizing the pixel-wise absolute difference between the sets H_α and H_γ associated to the real and fake frames, as follows:

$$\mathcal{M} = \frac{1}{N}\sum_{n=1}^{N}\left|H_\alpha^n - H_\gamma^n\right|, \tag{8}$$

$$\mathcal{M} = \frac{\mathcal{M} - min(\mathcal{M})}{max(\mathcal{M}) - min(\mathcal{M})}, \tag{9}$$

where N is the total number of feature maps comprising the latent sets, while H_α^n and H_γ^n are the n-th feature map for original and reconstructed images, respectively. Notice that during tests, the localizer L generates the heat map $\mathcal{M}$ only if the detector D classifies the frame as anomalous, considerably reducing the overall network computational cost. An example of heat map generation for anomaly localization is shown in Figure 4.



(**a**) (**b**) (**c**)

Figure 4. Example of heat map generation in an abnormal frame during the test phase. In (**a**), the frame depicting an abnormal item (i.e., the abandoned backpack); in (**b**) the heat map $\mathcal{M}$ produced by localizer model; in (**c**) the resulting localization.

Concluding, the two branches D and L are trained jointly to enhance both detection and localization performances. Therefore, the objective of the proposed two-branch neural network is to minimize the loss function defined as:

$$\mathcal{L} = \mathcal{L}_{adv} + \mathcal{L}_D + \mathcal{L}_L. \tag{10}$$

4. Results

This section reports the results of the proposed two-branch GAN-based neural network architecture on anomaly detection and localization tasks. Initially, it provides details about

the UMCD dataset, a public collection of challenging RGB video sequences acquired by UAVs at low altitudes and suitable for the proposed tasks. Afterward, it reports implementation details as well as quantitative and qualitative results.

4.1. Dataset

The UMCD [42] is a benchmark dataset for mosaicking and change detection systems, containing low-altitude RGB video streams ranging from 6 to 15 m, with speeds ranging from 2 up to 12 m/s, at different daily times, i.e., morning and afternoon. Specifically, all videos were acquired employing two different small-scale UAVs, a DJI Phantom 3 Advanced with a built-in camera each, and a custom home-made hexacopter having cameras with different spatial resolutions, ranging from 720 × 540 to 1920 × 1080 pixels per frame. Concerning the visual data, the UCMD dataset includes 50 aerial video sequences collected in different environments, i.e., urban, dirt, and countryside (CS). Among these, only a subset of 20 video sequences was devised for change detection and can be adapted to anomaly-related tasks. Indeed, they consist of pairs of videos where each couple shows a given path with and without specific elements, as shown in Figure 5, which can be targeted as anomalous. Details are reported in Section 5.1. Finally, the remaining 30 videos were excluded from the model training since they do not present anomalies for tests.

(**a**) (**b**) (**c**)

Figure 5. In (**a**–**c**), image samples for different paths in the UMCD dataset. In the top row, frames associated to the normal scene. In the bottom row, the same frames depicting abnormal elements highlighted in the red bounding box.

4.2. Implementation Details

The proposed two-branch neural network architecture was implemented using the Pytorch framework, and, following the same experimental protocol, all experiments were performed on a single GPU, i.e., one GeForce RTX 3080 with 16 GB of RAM. Specifically, the network was trained on each environment separately using the corresponding 10 videos without anomalies via a semi-supervised paradigm. The remaining sequences containing anomalous elements were used to evaluate the model detection and localization capabilities. Each input frame was resized to a dimension of 256 × 256 to ensure an affordable computational cost while preserving the visibility of small anomalies. Higher resolutions do not provide better performance other than requiring more computational resources. Regarding the training process, AdamW [73] was used as the optimizer with a learning rate set to 0.0002, an ϵ parameter of 1×10^{-8}, a weight decay set to 1×10^{-2}, and a first β_1 and second β_2 momentum initial decay rate of 0.5 and 0.999, respectively. Finally, the weight parameters adjusting the loss functions for both detector D and localizer L components were set to $w_r^D = w_r^L = 0.8$ and $w_e^D = w_e^L = 0.2$. The network was trained for 200 epochs on each environment since the training loss $\mathcal{L}$ becomes steady for all backgrounds within that number of epochs. For the proposed method evaluation, common metrics were used for anomaly detection and localization. In particular, the anomaly detection being a binary classification problem, the AUROC was used to evaluate the detection capability of anoma-

lous elements. In detail, the area under the ROC curve indicates the ability of the detector to distinguish between two classes, i.e., normal or abnormal. For the anomaly localization, instead, the SSIM was used to measure the model capability to capture spatial differences with anomaly-free scenarios, i.e., abnormal regions. Such an index is a perceptual metric measuring the ability of the localizer to reconstruct normal scenes given as input.

4.3. Anomaly Detection Results

This subsection reports the obtained anomaly detection results for each kind of environment. To this end, measuring the AUROC metric, the true positive rate (TPR) is the percentage of frames correctly detected as anomalous over the observed environment. Instead, the false positive rate (FPR) is the percentage of frames wrongly detected as anomalous over the same observed scene. Table 1 shows the detection performance obtained using only the localizer L in an eventually single-branch setting, achieving an average AUROC of 82.6%. Table 2, instead, reports results for each background with different configurations in the proposed two-branch architectural design, achieving an increased average AUROC of 97.2%. The Euclidean distance among latent vectors being associated with repaired and anomalous images used as anomaly score, a threshold value is required to perform the final normal or abnormal classification. Since the threshold depends on the background, the optimal value for each environment is automatically learned as described in Section 3.1 and reported in Table 3.

Table 1. Performance evaluation of localizer L obtained by changing its number of layers for the detection task. Reported scores refer to the AUROC metric.

	Path Seq. #	**3-Layers**	**4-Layers**	**5-Layers**
CS	Path #1	0.839	0.856	0.847
	Path #2	0.831	0.825	0.827
Dirt	Path #1	0.871	0.875	0.879
	Path #2	0.852	0.850	0.873
	Path #3	0.788	0.765	0.770
	Path #4	0.798	0.763	0.783
Urban	Path #1	0.796	0.779	0.784
	Path #2	0.870	0.876	0.880
	Path #3	0.739	0.720	0.730
	Path #4	0.873	0.870	0.860
	Avg AUROC	**0.826**	**0.818**	**0.823**

Table 2. Performance evaluation of detector D obtained by changing its number of layers. Reported scores refer to the AUROC metric.

	Path Seq. #	**3-Layers**	**4-Layers**	**5-Layers**
CS	Path #1	0.979	0.976	0.959
	Path #2	0.966	0.969	0.970
Dirt	Path #1	0.968	0.974	0.977
	Path #2	0.980	0.976	0.973
	Path #3	0.977	0.965	0.949
	Path #4	0.952	0.968	0.957
Urban	Path #1	0.973	0.984	0.952
	Path #2	0.976	0.982	0.987
	Path #3	0.936	0.945	0.938
	Path #4	0.983	0.979	0.980
	Avg AUROC	**0.969**	**0.972**	**0.964**

Table 3. Listof automatically learned distance thresholds indicating the presence of anomalies for each path sequence in the dataset.

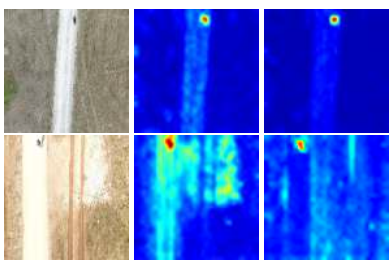
	Path Seq. #	Threshold
CS	Path #1	0.079
	Path #2	0.031
Dirt	Path #1	0.039
	Path #2	0.125
	Path #3	0.330
	Path #4	0.362
Urban	Path #1	0.332
	Path #2	0.275
	Path #3	0.219
	Path #4	0.011

4.4. Anomaly Localization Results

Regarding anomaly localization, quantitative and qualitative results are reported, also in this case, for each kind of environment. Table 4 summarizes the quantitative results for different configurations and each background using the SSIM metric, achieving an average of 95.7%. Concerning the qualitative results, the heat maps produced by the localizer L are reported. In detail, examples of $\mathcal{M}$ computed on a couple of anomalous frames with and without the support of residual blocks are illustrated in Figure 6b,c, respectively. Thus, examples of $\mathcal{M}$ computed with the support of residual blocks on different anomalous frames in various environments are depicted in Figure 7. By observing Figure 7b,e, the generated heat maps can highlight even multiple anomalous elements with different shapes, colors, sizes, and positions.

Table 4. Performance evaluation of localizer L obtained by changing its number of layers. Reported scores refer to the SSIM metric.

	Path Seq. #	3-Layers	4-Layers	5-Layers
CS	Path #1	0.956	0.948	0.951
	Path #2	0.947	0.950	0.946
Dirt	Path #1	0.955	0.960	0.959
	Path #2	0.962	0.949	0.953
	Path #3	0.960	0.968	0.963
	Path #4	0.956	0.952	0.944
Urban	Path #1	0.954	0.958	0.945
	Path #2	0.946	0.937	0.931
	Path #3	0.965	0.962	0.954
	Path #4	0.968	0.974	0.970
	Avg SSIM	**0.957**	**0.956**	**0.952**



(**a**) (**b**) (**c**)

Figure 6. In (**a**), image samples depicting abnormal conditions; in (**b**,**c**), the heat maps produced by the localizer without and with residual blocks, respectively.

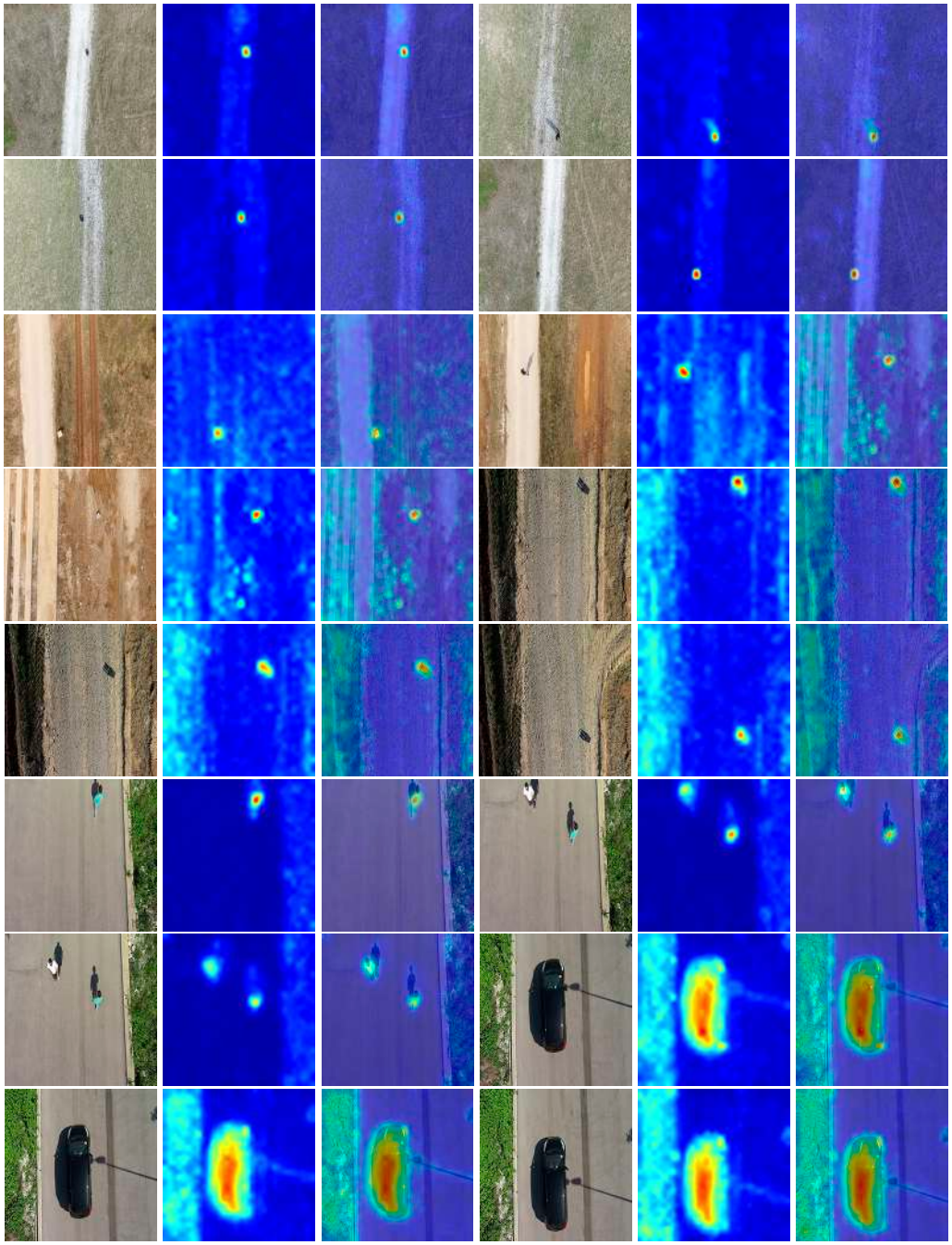


Figure 7. Examples of $\mathcal{M}$ (**b**,**e**) and resulting localization (**c**,**f**) computed on images (**a**,**d**) with different abnormal elements in multiple environments.

5. Discussion

This section initially motivates the choice of the UMCD dataset and the reported implementation details. Finally, it provides an exhaustive discussion on anomaly detection and localization results.

5.1. Dataset

The choice for using the UMCD dataset is related to its peculiarities, which make it suitable for the presented and addressed anomaly-related tasks. The reason is threefold. First, it is the only public dataset with low-altitude UAV-captured videos depicting the

same scene with and without specific elements along a given path that can be considered anomalies to detect. Such elements, varying in size, shape, color, and position, include tire, gas bottle, person, car, small box, big box, metal suitcase, suitcase, and bag. Second, videos are collected in three different environments comprising paths with diverse complexity, including multiple elements appearing simultaneously. Therefore, the method's robustness can also be tested in challenging conditions, e.g., background clutter or multiple targets. Third, it allows the simulation of a practical aerial video surveillance scenario because the objects included in the scenes can represent a security risk. For example, small and big boxes, suitcases, or bags contain dangerous objects, such as IEDs. At the same time, a car in a place where it is not allowed could represent a possible threat containing a bomb.

5.2. Implementation Details

All tested GAN models were trained using the AdamW optimizer because in the performed experiments it has shown a slight improvement in final results when compared to the Adam [74] version. Probably, this is due to the better model generalization provided by the former while also providing increased training speed. The two branches comprising the proposed architectures are jointly optimized for the entire loss specified in Equation (10) using an initial learning rate set to 0.0002 and the first momentum initial decay β_1 set to 0.5, leaving the suggested value for the second momentum instead, as in [75]. Concerning the learning rate, the experiments effectively highlighted that the recommended 0.001 was too high, resulting in the inability to reconstruct images. Moreover, changing β_1 avoids training oscillation and instability. Observe that all these settings are suggested for stabilizing the GAN-based networks training phase [62]. Finally, the weighting parameters adjusting the impact of individual losses on the overall loss function were set to $w_r^D = w_r^L = 0.8$ and $w_e^D = w_e^L = 0.2$, empirically.

5.3. Performance Evaluation

Extensive experiments were performed to evaluate the various method components. Specifically, this section discusses how the usage of two specialized branches, different numbers of convolutional layers, and the addition of residual blocks affect anomaly detection and localization tasks.

5.3.1. Single-Branch Architecture

Even if trained simultaneously, models D and L perform specific jobs and can work separately. However, only the localizer can inherently perform both detection and localization tasks thanks to its architectural design, whereas due to missing localization capability, the detector cannot be tested for a single-branch setting. Despite the acceptable performance obtained using only the localizer for both tasks, the results increase by jointly training D and L networks through the objective function $\mathcal{L}$ reported in Equation (10). The single-branch detection performance decrease is probably due to the capability of model L, whose structure is specialized for localization, in reconstructing abnormal elements in the scene, breaking the model D effective detection strategy leveraging anomaly correction.

5.3.2. Anomaly Detection

Regarding the anomaly detection task, the detector branch follows the anomaly repair strategy described in Section 3.1. In particular, once observing video frames depicting the normal scene during the training, detector D tries to fix any anomaly that might appear in the input frame. By repairing the latter during its reconstruction, in terms of Euclidean distance, the latent vector of an anomalous scene will be far away from the one representing the repaired image, which would be closer to the low-dimensional manifold of normal frames. This distance acts as an anomaly score and can be used for detecting anomalies in the input by employing a threshold value. The optimal threshold for each background defining the presence of anomalous elements is automatically learned since it is dependent on the environment, e.g., it is influenced by the ground complexity. An ablation study was

performed for the detector model evaluation varying the number of convolutional layers. Regarding the detection performance, significant results are achieved using only three convolutional layers. However, a slight increase in detection capability can be observed with four layers followed by a new decrease in adding a new layer, i.e., five layers in total. The latter is probably due to the noise introduced by too many operations inside the model. Therefore, extra layers were not tested. What is more, model D is robust across all the backgrounds independently from the number of layers, probably due to the cooperation of the two branches during the training, demonstrating the capability to detect any anomaly.

5.3.3. Anomaly Localization

Regarding the localization task, quantitative and qualitative evaluations were performed through experiments focused on anomaly localization and heat map generation. In order to reduce the overall network computational cost, during test time, the localizer branch is executed only when the detector predicts the presence of an anomaly within the processed video frame. Specifically, model L produces good quality heat maps containing spatial information on the observed scene employed to highlight the most meaningful areas of the video to monitor for noticing the anomalies, as described in Section 3.2. Concerning the quantitative results, like for detector D, an ablation study to assess the proper number of convolutional layers was also performed for the localization via the SSIM metric. In this case, a higher structural similarity index indicates the great localizer L capability to reconstruct the original frame and localize anomalies inside the depicted scene. As can be noticed, remarkable results are obtained with only three layers, while there is no significant improvement or deterioration in increasing the number of convolutional layers to four or five. Therefore, three layers can be the better option for anomaly localization, reducing the number of operations performed inside the model while achieving good performance. Similar to the detector, model L is robust across the environments, confirming that the localizer structure extracts a good quality latent set of feature maps that correctly characterize the input allowing its reconstruction. As can be observed in Figure 6c, the presence of the residual blocks improves the heat map quality that is more accurate and less noisy.

6. Conclusions

In this paper, a novel two-branch GAN-based method for RGB aerial video streams acquired by UAVs at low altitude for surveillance purposes is presented. In particular, we have shown that even in complex operational scenarios, it is possible to use an AD paradigm to detect and localize both common elements (e.g., vehicles, persons) as well as very small objects (e.g., gas bottles, boxes, suitcases) that can represent a reason for danger or attention. Even if we have analyzed the current state of the art related to the databases acquired by small-scale UAVs, we have observed that only the UMCD dataset has, all together, the distinctive characteristics that make it suitable for our purposes. In fact, it is the only dataset whose video sequences are acquired at very low altitudes (i.e., between 6 and 15 m), with different parameters (e.g., speed, height) on different types of ground (i.e., urban, dirt, and countryside), having natural and artificial structures (e.g., trees, buildings) as well as specific objects different in color, size, position, and shape. In addition, the UMCD dataset is the only one in the literature that presents the same paths acquired twice, with and without anomalous objects. The performed tests measured by using the common metrics related to this application area have shown outstanding results in terms of AUROC (97.2%) and SSIM (95.7%), showing the ability of the system in performing the binary classification and the similarity task, respectively. We really hope that the proposed work and the related dataset can become a comparative baseline for future applications in this field.

Author Contributions: Conceptualization, D.A., M.C., A.F., G.L.F. and D.P.; methodology, D.A., M.C., L.C., G.L.F., M.M. and D.P.; formal analysis, L.C., G.L.F. and M.M.; investigation, I.C., M.C., L.C., A.D., A.F., G.L.F., R.L., M.M., A.M. and D.P.; software, I.C., M.C., A.D., A.F., R.L., A.M. and D.P.; validation, I.C., M.C., L.C., A.D., A.F., G.L.F., R.L., M.M. and A.M.; data curation, I.C., M.C., A.D., A.F., R.L., A.M.

and D.P.; writing—original draft preparation, D.A., I.C., M.C., A.D., A.F., R.L., M.M., A.M. and D.P.; writing—review and editing, D.A., I.C., M.C., A.D., A.F., R.L., M.M., A.M. and D.P.; supervision, L.C. and G.L.F. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the MIUR under grant "Departments of Excellence 2018–2022" of the Department of Computer Science of Sapienza University, the "Smart unmannEd AeRial vehiCles for Human likE monitoRing (SEARCHER)" project of the Italian Ministry of Defence (CIG: Z84333EA0D), and the ERC Starting Grant no. 802554 (SPECGEO).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Huang, C.R.; Huang, W.Y.; Liao, Y.S.; Lee, C.C.; Yeh, Y.W. A Content-Adaptive Resizing Framework for Boosting Computation Speed of Background Modeling Methods. *IEEE Trans. Syst. Man Cybern. Syst.* **2022**, *52*, 1192–1204. [CrossRef]
2. Wang, H.; Lv, X.; Zhang, K.; Guo, B. Building Change Detection Based on 3D Co-Segmentation Using Satellite Stereo Imagery. *Remote Sens.* **2022**, *14*, 628. [CrossRef]
3. Avola, D.; Bernardi, M.; Cinque, L.; Foresti, G.L.; Massaroni, C. Adaptive Bootstrapping Management by Keypoint Clustering for Background Initialization. *Pattern Recognit. Lett.* **2017**, *100*, 110–116. [CrossRef]
4. Yang, L.; Cheng, H.; Su, J.; Li, X. Pixel-to-Model Distance for Robust Background Reconstruction. *IEEE Trans. Circ. Syst. Video Technol.* **2016**, *26*, 903–916. [CrossRef]
5. Zhang, X.; Huang, T.; Tian, Y.; Gao, W. Background-Modeling-Based Adaptive Prediction for Surveillance Video Coding. *IEEE Trans. Image Process.* **2014**, *23*, 769–784. [CrossRef]
6. Jing, W.; Zhu, S.; Kang, P.; Wang, J.; Cui, S.; Chen, G.; Song, H. Remote Sensing Change Detection Based on Unsupervised Multi-Attention Slow Feature Analysis. *Remote Sens.* **2022**, *14*, 2834. [CrossRef]
7. Gong, H.; Mu, T.; Li, Q.; Dai, H.; Li, C.; He, Z.; Wang, W.; Han, F.; Tuniyazi, A.; Li, H.; et al. Swin-Transformer-Enabled YOLOv5 with Attention Mechanism for Small Object Detection on Satellite Images. *Remote Sens.* **2022**, *14*, 2861. [CrossRef]
8. Avola, D.; Foresti, G.L.; Cinque, L.; Massaroni, C.; Vitale, G.; Lombardi, L. A Multipurpose Autonomous Robot for Target Recognition in Unknown Environments. In Proceedings of the 14th IEEE International Conference on Industrial Informatics (INDIN), Poitiers, France, 19–21 July 2016; pp. 766–771.
9. Pan, X.; Tang, F.; Dong, W.; Gu, Y.; Song, Z.; Meng, Y.; Xu, P.; Deussen, O.; Xu, C. Self-Supervised Feature Augmentation for Large Image Object Detection. *IEEE Trans. Image Process.* **2020**, *29*, 6745–6758. [CrossRef]
10. del Blanco, C.R.; Jaureguizar, F.; Garcia, N. An Efficient Multiple Object Detection and Tracking Framework for Automatic Counting and Video Surveillance Applications. *IEEE Trans. Consum. Electron.* **2012**, *58*, 857–862. [CrossRef]
11. He, C.; Zhang, J.; Yao, J.; Zhuo, L.; Tian, Q. Meta-Learning Paradigm and CosAttn for Streamer Action Recognition in Live Video. *IEEE Signal Process. Lett.* **2022**, *29*, 1097–1101. [CrossRef]
12. Liu, T.; Ma, Y.; Yang, W.; Ji, W.; Wang, R.; Jiang, P. Spatial-Temporal Interaction Learning Based Two-Stream Network for Action Recognition. *Inf. Sci.* **2022**, *606*, 864–876. [CrossRef]
13. Meng, Q.; Zhu, H.; Zhang, W.; Piao, X.; Zhang, A. Action Recognition Using Form and Motion Modalities. *ACM Trans. Multimed. Comput. Commun. Appl.* **2020**, *16*, 1–16. [CrossRef]
14. Avola, D.; Cinque, L.; Foresti, G.L.; Pannone, D. Automatic Deception Detection in RGB Videos Using Facial Action Units. In Proceedings of the 13th International Conference on Distributed Smart Cameras (ICDSC), Trento, Italy, 9–11 September 2019; pp. 1–6.
15. Zhao, Q.; Zhang, B.; Lyu, S.; Zhang, H.; Sun, D.; Li, G.; Feng, W. A CNN-SIFT Hybrid Pedestrian Navigation Method Based on First-Person Vision. *Remote Sens.* **2018**, *10*, 1229. [CrossRef]
16. Maji, B.; Swain, M.; Mustaqeem. Advanced Fusion-Based Speech Emotion Recognition System Using a Dual-Attention Mechanism with Conv-Caps and Bi-GRU Features. *Electronics* **2022**, *11*, 1328. [CrossRef]
17. Liao, Y.; Vakanski, A.; Xian, M. A Deep Learning Framework for Assessing Physical Rehabilitation Exercises. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2020**, *28*, 468–477. [CrossRef] [PubMed]
18. Vamsikrishna, K.M.; Dogra, D.P.; Desarkar, M.S. Computer-Vision-Assisted Palm Rehabilitation With Supervised Learning. *IEEE Trans. Biomed. Eng.* **2016**, *63*, 991–1001. [CrossRef]
19. Petracca, A.; Carrieri, M.; Avola, D.; Basso Moro, S.; Brigadoi, S.; Lancia, S.; Spezialetti, M.; Ferrari, M.; Quaresima, V.; Placidi, G. A Virtual Ball Task Driven by Forearm Movements for Neuro-Rehabilitation. In Proceedings of the International Conference on Virtual Rehabilitation (ICVR), Valencia, Spain, 9–12 June 2015; pp. 162–163.
20. Du, D.; Han, X.; Fu, H.; Wu, F.; Yu, Y.; Cui, S.; Liu, L. *SAniHead*: Sketching Animal-Like 3D Character Heads Using a View-Surface Collaborative Mesh Generative Network. *IEEE Trans. Vis. Comput. Graph.* **2022**, *28*, 2415–2429. [CrossRef] [PubMed]
21. Jackson, B.; Keefe, D.F. Lift-Off: Using Reference Imagery and Freehand Sketching to Create 3D Models in VR. *IEEE Trans. Vis. Comput. Graph.* **2016**, *22*, 1442–1451. [CrossRef] [PubMed]
22. Avola, D.; Caschera, M.C.; Ferri, F.; Grifoni, P. Ambiguities in Sketch-Based Interfaces. In Proceedings of the 40th Annual Hawaii International Conference on System Sciences (HICSS), Waikoloa, HI, USA, 3–6 January 2007; p. 290b.

23. Messina, G.; Modica, G. Applications of UAV Thermal Imagery in Precision Agriculture: State of the Art and Future Research Outlook. *Remote Sens.* **2020**, *12*, 1491. [CrossRef]
24. AdÃ£o, T.; HruÅ¡ka, J.; PÃ¡dua, L.; Bessa, J.; Peres, E.; Morais, R.; Sousa, J.J. Hyperspectral Imaging: A Review on UAV-Based Sensors, Data Processing and Applications for Agriculture and Forestry. *Remote Sens.* **2017**, *9*, 1110. [CrossRef]
25. Marusic, Z.; Zelenika, D.; Marusic, T.; Gotovac, S. Visual Search on Aerial Imagery as Support for Finding Lost Persons. In Proceedings of the 8th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro, 10–14 June 2019; pp. 1–4.
26. Scherer, J.; Yahyanejad, S.; Hayat, S.; Yanmaz, E.; Andre, T.; Khan, A.; Vukadinovic, V.; Bettstetter, C.; Hellwagner, H.; Rinner, B. An Autonomous Multi-UAV System for Search and Rescue. In Proceedings of the Workshop on Micro Aerial Vehicle Networks, Systems, and Applications for Civilian Use (DroNet), Florence, Italy, 18 May 2015; pp. 33–38.
27. Ul Ain Tahir, H.; Waqar, A.; Khalid, S.; Usman, S.M. Wildfire Detection in Aerial Images Using Deep Learning. In Proceedings of the 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2), Rawalpindi, Pakistan, 24–26 May 2022; pp. 1–7.
28. Jiao, Z.; Zhang, Y.; Mu, L.; Xin, J.; Jiao, S.; Liu, H.; Liu, D. A YOLOv3-based Learning Strategy for Real-time UAV-based Forest Fire Detection. In Proceedings of the Chinese Control and Decision Conference (CCDC), Hefei, China, 22–24 August 2020; pp. 4963–4967.
29. Xiao, J.; Zhang, S.; Dai, Y.; Jiang, Z.; Yi, B.; Xu, C. Multiclass Object Detection in UAV Images Based on Rotation Region Network. *IEEE J. Miniat. Air Space Syst.* **2020**, *1*, 188–196. [CrossRef]
30. Zhang, R.; Shao, Z.; Huang, X.; Wang, J.; Li, D. Object Detection in UAV Images via Global Density Fused Convolutional Network. *Remote Sens.* **2020**, *12*, 3140. [CrossRef]
31. Wang, S.; Han, Y.; Chen, J.; Zhang, Z.; Wang, G.; Du, N. A Deep-Learning-Based Sea Search and Rescue Algorithm by UAV Remote Sensing. In Proceedings of the IEEE CSAA Guidance, Navigation and Control Conference (CGNCC), Xiamen, China, 10–12 August 2018; pp. 1–5.
32. Avola, D.; Cinque, L.; Di Mambro, A.; Diko, A.; Fagioli, A.; Foresti, G.L.; Marini, M.R.; Mecca, A.; Pannone, D. Low-Altitude Aerial Video Surveillance via One-Class SVM Anomaly Detection from Textural Features in UAV Images. *Information* **2022**, *13*, 2. [CrossRef]
33. Avola, D.; Pannone, D. MAGI: Multistream Aerial Segmentation of Ground Images with Small-Scale Drones. *Drones* **2021**, *5*, 111. [CrossRef]
34. Avola, D.; Cinque, L.; Fagioli, A.; Foresti, G.L.; Pannone, D.; Piciarelli, C. Automatic Estimation of Optimal UAV Flight Parameters for Real-Time Wide Areas Monitoring. *Multimed. Tools Appl.* **2021**, *80*, 25009–25031. [CrossRef]
35. Avola, D.; Cinque, L.; Diko, A.; Fagioli, A.; Foresti, G.L.; Mecca, A.; Pannone, D.; Piciarelli, C. MS-Faster R-CNN: Multi-Stream Backbone for Improved Faster R-CNN Object Detection and Aerial Tracking from UAV Images. *Remote Sens.* **2021**, *13*, 1670. [CrossRef]
36. Avola, D.; Foresti, G.L.; Martinel, N.; Micheloni, C.; Pannone, D.; Piciarelli, C. Real-Time Incremental and Geo-Referenced Mosaicking by Small-Scale UAVs. In Proceedings of the International Conference on Image Analysis and Processing (ICIAP), Catania, Italy, 11–15 September 2017; pp. 694–705.
37. Avola, D.; Cinque, L.; Fagioli, A.; Foresti, G.L.; Massaroni, C.; Pannone, D. Feature-Based SLAM Algorithm for Small Scale UAV with Nadir View. In Proceedings of the International Conference on Image Analysis and Processing (ICIAP), Trento, Italy, 9–13 September 2019; pp. 457–467.
38. Diez, Y.; Kentsch, S.; Fukuda, M.; Caceres, M.L.L.; Moritake, K.; Cabezas, M. Deep Learning in Forestry Using UAV-Acquired RGB Data: A Practical Review. *Remote Sens.* **2021**, *13*, 2837. [CrossRef]
39. Bozcan , I.; Kayacan, E. UAV-AdNet: Unsupervised Anomaly Detection using Deep Neural Networks for Aerial Surveillance. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 24 October 2020–24 January 2021; pp. 1158–1164.
40. Chriki, A.; Touati, H.; Snoussi, H.; Kamoun, F. UAV-based Surveillance System: An Anomaly Detection Approach. In Proceedings of the IEEE Symposium on Computers and Communications (ISCC), Rennes, France, 7–10 July 2020; pp. 1–6.
41. Martin, R.A.; Blackburn, L.; Pulsipher, J.; Franke, K.; Hedengren, J.D. Potential Benefits of Combining Anomaly Detection with View Planning for UAV Infrastructure Modeling. *Remote Sens.* **2017**, *9*, 434. [CrossRef]
42. Avola, D.; Cinque, L.; Foresti, G.L.; Martinel, N.; Pannone, D.; Piciarelli, C. A UAV Video Dataset for Mosaicking and Change Detection From Low-Altitude Flights. *IEEE Trans. Syst. Man Cybern. Syst.* **2020**, *50*, 2139–2149. [CrossRef]
43. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv.* **2009**, *41*, 1–58. [CrossRef]
44. Ramachandra, B.; Jones, M.J.; Vatsavai, R.R. A Survey of Single-Scene Video Anomaly Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 2293–2312. [CrossRef] [PubMed]
45. Nayak, R.; Pati, U.C.; Das, S.K. A Comprehensive Review on Deep Learning-Based Methods for Video Anomaly Detection. *Image Vis. Comput.* **2021**, *106*, 1–19. [CrossRef]
46. Pang, G.; Shen, C.; Cao, L.; Hengel, A.V.D. Deep Learning for Anomaly Detection: A Review. *ACM Comput. Surv.* **2021**, *54*, 1–38. [CrossRef]
47. Hamdi, S.; Bouindour, S.; Snoussi, H.; Wang, T.; Abid, M. End-to-End Deep One-Class Learning for Anomaly Detection in UAV Video Stream. *J. Imaging* **2021**, *7*, 90. [CrossRef]

48. Chan, A.; Vasconcelos, N. UCSD Pedestrian Dataset. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 909–926. [CrossRef]
49. Bonetto, M.; Korshunov, P.; Ramponi, G.; Ebrahimi, T. Privacy in Mini-Drone Based Video Surveillance. In Proceedings of the 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia, 4–8 May 2015; pp. 1–6.
50. Chriki, A.; Touati, H.; Snoussi, H.; Kamoun, F. Deep Learning and Handcrafted Features for One-Class Anomaly Detection in UAV Video. *Multimed. Tools Appl.* **2021**, *80*, 2599–2620. [CrossRef]
51. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.
52. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 886–893.
53. Pearson, K. LIII. On Lines and Planes of Closest Fit to Systems of Points in Space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1901**, *2*, 559–572. [CrossRef]
54. Klaser, A.; Marszałek, M.; Schmid, C. A Spatio-Temporal Descriptor Based on 3D-Gradients. In Proceedings of the 19th British Machine Vision Conference (BMVC), Leeds, UK, 1–4 September 2008 ; pp. 1–10.
55. Boser, B.E.; Guyon, I.M.; Vapnik, V.N. A Training Algorithm for Optimal Margin Classifiers. In Proceedings of the 5th Annual Workshop on Computational Learning Theory, Pittsburgh, PA, USA, 27–29 July 1992; pp. 144–152.
56. Jin, P.; Mou, L.; Xia, G.S.; Zhu, X.X. Anomaly Detection in Aerial Videos Via Future Frame Prediction Networks. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Brussels, Belgium, 11–16 July 2021; pp. 8237–8240.
57. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
58. Haralick, R.M.; Shanmugam, K.; Dinstein, I.H. Textural Features for Image Classification. *IEEE Trans. Syst. Man Cybern.* **1973**, *SMC-3*, 610–621. [CrossRef]
59. Haralick, R.M. Statistical and Structural Approaches to Texture. *Proc. IEEE* **1979**, *67*, 786–804. [CrossRef]
60. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the International Conference on Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 8–11 December 2014; pp. 2672–2680.
61. Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. *arXiv* **2014**, arXiv:abs/1411.1784.
62. Radford , A.; Metz, L.; Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In Proceedings of the International Conference on Learning Representations (ICLR), San Juan, Puerto Rico, 2–4 May 2016; pp. 1–20.
63. Avola, D.; Cascio, M.; Cinque, L.; Fagioli, A.; Foresti, G.L. Human Silhouette and Skeleton Video Synthesis Through Wi-Fi Signals. *Int. J. Neural Syst.* **2022**, *32*, 1–20. [CrossRef] [PubMed]
64. Akcay, S.; Atapour-Abarghouei, A.; Breckon, T.P. Ganomaly: Semi-Supervised Anomaly Detection via Adversarial Training. In Proceedings of the Asian Conference on Computer Vision (ACCV), Perth, Australia, 2–6 December 2018; pp. 622–637.
65. Chen, D.; Yue, L.; Chang, X.; Xu, M.; Jia, T. NM-GAN: Noise-Modulated Generative Adversarial Network for Video Anomaly Detection. *Pattern Recognit.* **2021**, *116*, 107969. [CrossRef]
66. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2015; pp. 448–456.
67. Maas, A.; Hannun, A.; Ng, A. Rectifier Nonlinearities Improve Neural Network Acoustic Models. In Proceedings of the International Conference on Machine Learning (ICML), Atlanta, GA, USA, 16–21 June 2013; pp. 1–6.
68. Youden, W.J. Index for Rating Diagnostic Tests. *Cancer* **1950**, *3*, 32–35. [CrossRef]
69. Carrara, F.; Amato, G.; Brombin, L.; Falchi, F.; Gennaro, C. Combining GANs and AutoEncoders for Efficient Anomaly Detection. In Proceedings of the International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 3939–3946.
70. Schlegl, T.; Seeböck, P.; Waldstein, S.M.; Langs, G.; Schmidt-Erfurth, U. f-AnoGAN: Fast Unsupervised Anomaly Detection with Generative Adversarial Networks. *Med. Image Anal.* **2019**, *54*, 30–44. [CrossRef]
71. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
72. Dong, L.F.; Gan, Y.Z.; Mao, X.L.; Yang, Y.B.; Shen, C. Learning Deep Representations Using Convolutional Auto-Encoders with Symmetric Skip Connections. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 3006–3010.
73. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019; pp. 1–19.
74. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015; pp. 1–15.
75. Avola, D.; Cascio, M.; Cinque, L.; Fagioli, A.; Foresti, G.L.; Marini, M.R.; Rossi, F. Real-time deep learning method for automated detection and localization of structural defects in manufactured products. *Comput. Ind. Eng.* **2022**, *172*, 108512. [CrossRef]