Transforming Temporal Embeddings to Keypoint Heatmaps for Detection of Tiny Vehicles in Wide Area Motion Imagery (WAMI) Sequences

Farhood Negin [1,2] Mohsen Tabejamaat [1] Renaud Fraisse [3] Francois Bremond [1,2]
[1] Inria [2] Université Côte d'azur [3] Airbus Defence & Space
{farhood.negin, mohsen.tabejamaat, francois.bremond}@inria.fr renaud.fraisse@airbus.com

Abstract

Nowadays, due to its many applications, objects detection in wide area motion imagery (WAMI) sequences has received a lot of attention. Unlike natural images, object detection in WAMI faces unique challenges. Lack of appearance information due to the small size of objects makes object detection difficult for conventional methods. In addition, pixel noise, registration errors, sparse or densely populated objects, brings on pronounced artifacts which amplifies the difficulty of detection. This paper aims to address object detection problem in the presence of these issues by considering objects as keypoints in the relevant background and proposes a spatiotemporal anchor-free detector for tiny vehicles in WAMI images. Instead of background subtraction, a region of interest network refines large search space of sequences to indicates object clusters. For further investigation, clusters are encoded by a codebook which is learned through an unsupervised encoder-decoder network. To accurately generate the detections, a Transformer network is trained on cluster embeddings using ground-truth heatmaps that are described by Gaussian distribution rather than hard label annotation. The network is trained with a redesigned version of Focal loss comprising a shape prior regularizer which help the generated heatmaps to conform to the shape of the keypoints. Extensive experiments on WPAFB dataset demonstrate the high capability of our method for the detection of small vehicles where it achieves competitive performance when compared to the state-of-the-art.

1. INTRODUCTION

Today, with the advent of high-altitude airborne and space cameras, long-term WAMI video recording has become possible, which has many applications in the military and civilian fields [2, 13, 17, 21, 22, 24, 32, 35]. For example, in urban planning, it can help control traffic and analyze drivers' behavior, and in security monitoring scenarios, it can help to identify dangerous and abnormal behaviors. In all of these applications, object detection has been one of the major techniques for achieving quantitative surveillance from videos. Numerous detection methods have been extensively studied and applied to ground-based surveillance systems in which videos are usually recorded by a fixed camera with a relatively low height from the ground and usually of high quality. Satellite remote sensing videos, on the other hand, are often recorded with low resolution and a non-stationary camera platform, and the recorded data usually contains geometrical deformations. Therefore, surveillance systems for processing such data face different challenges than traditional methods applied on ground-based systems and natural images. At a ground sample distance of about 1 meter, the target moving vehicles in current satellite videos are spatially recorded at approximately 5 to 20 pixels. Therefore, a vehicle in these images, which are usually captured in gray-level, is consisted of only a few pixels without color, shape and distinctive texture information. This makes the detection task very challenging in particular for the anchor-based methods at which the predefined anchors need to cover the objects very well. For tiny objects, a few number of pixels can produce large errors as it is not simple to determine whether a pixel near an

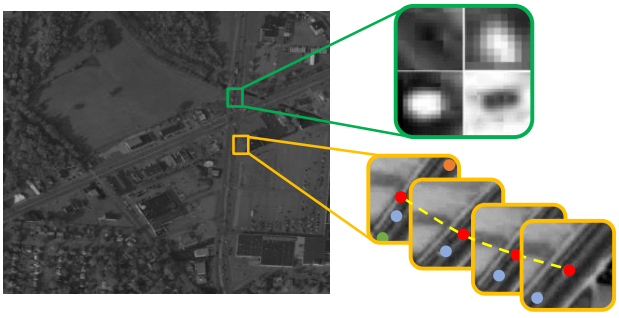

Figure 1. Sometimes small vehicles are indistinguishable from noise: in the green box, the top row shows real vehicles. The bottom row is just noise. Effective modeling of temporal features is critical for more accurate detection of moving objects (yellow boxes).

anchor belongs to an object or not.

Lack of appearance information makes motion the most powerful feature to identify these objects. Nonetheless this feature is often challenged in detection task due to background movement. In addition, since satellite camera systems often consist of several separate sensors, this can create artifacts such as lighting discontinuities and shearing. Accordingly, moving vehicles in the scene can be confused with pixel noise and, in general, with noise patterns from complex moving backgrounds, which creates problems for detection (Figure 1). Furthermore, objects in high resolution WAMI images are often sparse and located only in certain areas, such as roads and around buildings. This adds to the complexity of detecting objects in the whole image, which produces a lot of false alarms and thereby reduces the efficiency of the system.

In this paper, we propose a novel spatiotemporal anchor-free strategy without requiring any post-processing to address the challenges of object detection task that are specific to WAMI sequences. To avoid problems regarding anchor-based methods, such as designing anchors and tuning their hyperparameters, we consider tiny vehicle detection as a keypoint detection problem. At first, the framework learns to detect regions of images containing sparse or dense clusters of objects through a lightweight convolutional network and reduces the wide search space of the large-scale images using their corresponding Gaussian heatmaps. Varying from conventional CNN-based representation, the model exploits spatiotemporal information of the sequences using the Vector Quantized Variational Autoencoder (VQ-VAE) technique to achieve an efficient discrete video-level representation. Unlike most of the current solutions that benefit only from limited spatial information and ignore the temporal dependency of targets or utilize inefficient sliding windows, our vector quantized embedding allows an efficient use of both appearance and motion information to surpass this limitation in an unsupervised manner. This way, we encourage the framework to encode the spatial characteristic of the samples with motion dynamics which leads to a further rich representation describing spatiotemporal evolution of the data. For precise detection of the objects, the embeddings are then used to train a Transformer network that takes clip-level embeddings as input and generates frame-level detection by estimating center of the objects. Therefore, our contributions can be summarized as follows:

- A region of interest network based on Gaussian heatmap of sequences is trained that significantly reduces the large search space of WAMI images.
- knowing the inadequate efficiency of standard feature representations and considering the scarcity of appearance features in WAMI images, we introduce a novel strategy that benefits from a combined spatiotemporal feature in an unsupervised manner.
- Unlike the current strategies, our method does not require any inefficient and error-prone background subtraction operations, which allows for effective detection of moving as well as stationary objects in the scene.
- To generate precise detection masks, we adapt a loss function with a shape prior regularizer to compensate for the unbalanced distribution of the objects presented in sparse sequences and also to achieve pixel-level segmentation heatmaps, which in return helps in improving accuracy of the detection.
- The proposed method achieves competitive performance in the detection task compared to the state-of-the-art.

2. Related work

It has been practically and theoretically proven that it is more difficult to detect smaller objects, especially in WAMI images. It has been shown in practice [12] that there is a big discrepancy between the performance of models on small objects and their performance on large objects. Theoretically, it can be explained that since the models use the information that passes through the model, they gradually form and learn the features required for detection. The lower the information flow (in the case of WAMI, only a few pixels), the weaker and less distinctive the features learned. Therefore, the loss function, which works on the basis of pixel computations, will not have enough signal to learn during backpropagation. To address these issues, object detection methods generally use spatial information, temporal information, or both.

2.1. Background subtraction methods for WAMI

The most straightforward approach to use temporal information is to calculate frame differences between two or a higher number of consecutive frames [13, 24, 32]. Despite simplicity this method is associated with the problem of aperture for slow objects where the inner pixels of objects are recognized as the background, or the ghost effect where the pixels behind moving objects are recognized as the foreground. Background subtraction [14, 17, 26] takes a more holistic view of the problem. Considering the detection of moving objects as a special case of foreground segmentation, it creates an explicit background model that recognizes foreground objects by subtraction of the current input image from this background. Parallax effect, difficulty in detection of the stationary targets, and change in global illumination that negatively affect modeling of the

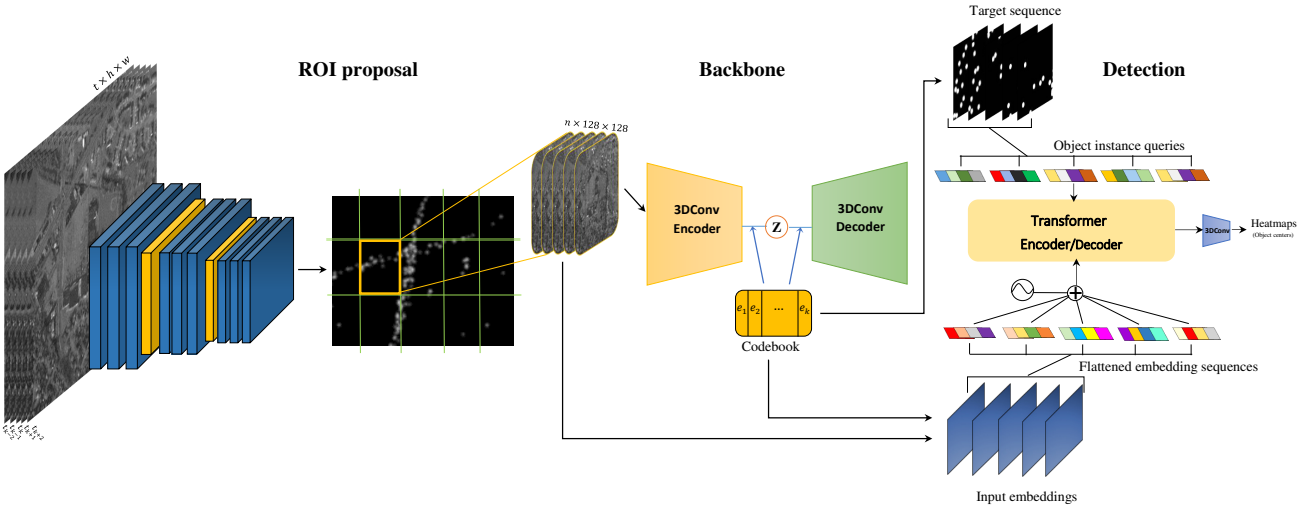

Figure 2. shows the overall architecture of our method which is comprised of three main modules: ROI proposal generator which is a CNN network that identifies areas that are most likely to contain objects. Backbone module that computes representative features of the generated proposals by training a VQ-VAE model. The detection module that uses feature embeddings to generate the detection heatmaps.

background are considered as the main issues of these methods.

2.2. Spatiotemporal networks

Region proposal networks (RPN) such as Faster-RCNN, which do not use temporal information and only detect objects in a single frame, have been shown [16] to fail in WAMI images. WAMI images are not only large and can not easily be used as input to such networks, but even splitting them into grids does not help because it does not conform to the principle of performance on which RPN networks are developed. It requires to design more anchors and tune their parameters that can not be generalized from other datasets. In addition, downsampling may not be a good choice as there is almost no reliable information left after downsampling due to the small size of the target objects (Which are usually annotated with only one point in the ground-truth). The benefits of using spatiotemporal information in various tasks have been exploited in recent years. Various network architectures [9, 10] have been tried in order to detect objects in videos. Some methods [18, 30] rely on post-processing of temporal information to make object detection results more coherent and consistent. These methods first use an off-the-shelf detector to find objects in each of the frames and then try to associate the results together. Some other work [6, 9, 10, 29, 33] takes advantage of the aggregation of spatiotemporal features. In particular, by designing special operators, they mainly improve the features of the current frame by aggregation of the features of adjacent frames (or the whole clip) and ultimately increase the overall detection performance. Very recently, Transformers [4, 7, 11, 41] raised great attention on various vision tasks and some methods use them to design object detection systems that achieve on par performance with state-of-the-art CNN-based detectors. However, most of these methods require complex post-processing stages to link the same object throughout the video to form tubelets that usually have a long convergence time. Consequently, only a handful of studies [16, 40] have used these methods on WAMI sequences.

3. Method

Figure 2 shows the overall architecture of our method. It is comprised of three main modules. First, to limit the search space on wide input images, a CNN network identifies areas of interest that are most likely to contain objects. In the second module an unsupervised model of representative features is learned from the generated proposals. The feature embeddings are then used as input to a Transformer in order to generate the outputs of the framework which are the generated heatmaps and the detection centers.

3.1. Regions of interest (ROI) proposals by regression heatmaps

Many CNN-based methods directly detect keypoint coordinates through regression [37, 38] or use anchors [3, 34] to extract deep features of candidate regions. Regression-based detectors require dense layers with large network parameters to learn image-to-coordinate mapping. This results in highly nonlinear models that are difficult to learn. Anchor-based methods, on the other hand, require anchor design in advance, and given the small number of object pixels in WAMI images, only the very precise part of the anchor can be used as a positive example. At this stage, finding the exact coordinates of objects is not our goal. The goal is to learn a lightweight network to find areas that po-

tentially contain vehicles, thereby reducing the large search space of WAMI images. Some methods do this by using overlaid road maps [36] where in addition to requiring extra information about images, they are not effective in scenarios that require the study of off-road objects. Our model adopts an anchor-free approach using heatmap-based proposal generation. Similar to [16, 27], proposals are generated by regression on heatmap images. The ground-truth information of objects in a specific pixel position are probabilistically encoded into heatmaps. The proposal generator takes a stack of five consecutive images as input and, after processing, generates a heatmap in which the approximate location of an object or group of objects is determined with Gaussian peaks. Input images are aligned based on a feature-based method. Detected FAST features [23] are used to compute descriptors. By matching descriptors with their corresponding pairs between two consecutive frames, a homography is estimated using the RANSAC algorithm [8] to warp them into a common reference. By having N the total number of the object centers in the ground-truth $O_i, i = \{1, \ldots, N\}$ with center coordinate of $x_i \in \mathbb{R}$, the d dimensional heatmap $H(x, \sigma) : \mathbb{R}^d \rightarrow \mathbb{R}$ for all pixels is obtained by

$$H(x, \sigma) = \sum_{i=1}^{N} \frac{1}{(2\pi)^{d/2}\sigma_i^d} e^{-\frac{||x-x_i||_2^2}{2\sigma_i^2}} \quad (1)$$

with high values around x_i which decreases as the pixel goes farther away from a ground-truth keypoint center. Considering that the objective at this stage is to detect clusters of objects, rather than accurate detection, the Euclidean loss between the network outputs and the ground-truth heatmap is used for loss calculations.

3.2. Learning latent encodings with VQ-VAE

The previous step gives us the regions of the input frames that may contain one or multiple objects, exempting us from an exhaustive search of the entire image. These smaller regions (128×128) are then searched in more details to determine the exact location of the objects. VQ-VAE [19] is utilized to represent these regions. VQ-VAE is a model that learns to compress high-dimensional data into a discrete hidden space in an unsupervised manner, which is potentially more suitable for many complex reasoning and predictive learning methods. In addition, it avoids the problem of "posterior collapse", which is one of the main problems of many VAE models due to the creation of powerful decoders that ignore the latents. In the case of videos, the goal is to learn a set of discrete latent representations from the raw pixels of video frames. To learn these discrete representations, instead of a one by one mapping between input data and latent variables, VQ-VAE learns a codebook and then forces the latents to be mapped to the nearest vector of the learned codebook. Given a video sample V, the encoder $E(V)$: $q_\theta(z|V)$ by passing through 3D convolutional layers (parameterized by θ), encodes the video frames into a series of latent continuous representations $\hat{z} \in \{1, \ldots, K\}$. Then, to obtain posterior distributions ($q_\theta(z|V)$), the continuous latents are discretized using a codebook $C = [e_1, e_2, \ldots, e_k]$ (which is a collection of K continuous vectors), by looking up the nearest element ($L2$ distance) in the codebook to the latent ($\hat{z}$):

$$q_\theta(z = c|V) = \begin{cases} 1 & \text{for c} = argmin_i ||\hat{z} - e_i||_2 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Then, the decoder $D(e)$: $q_\phi(V|e_c)$ uses the codebook embeddings of the input to generate the original input $\hat{V}$. VQ-VAE is trained by the following objective function:

$$\mathcal{L} = \|V - D(e)\|_2^2 + \|sg[E(V)] - e\|_2^2 + \beta\|sg[e] - E(V)\|_2^2 \quad (3)$$

where $sg(.)$ is the stop-gradient operator and the function can be summarized as a three-terms loss:

$$\mathcal{L} = \mathcal{L}_{reconstruct} + \mathcal{L}_{codebook} + \mathcal{L}_{commitment} \quad (4)$$

The reconstruction loss is a log-likelihood function that encourages the VQ-VAE to learn accurate representations by optimizing encoder and decoder parameters. Codebook loss's aim is to optimize codebook vectors in order to make embeddings closer to their latent encoder outputs. Third term with β scaling hyperparameter, prevents the encoder from fluctuating representations and makes it to bind to the representations which are closer to the codebook.

3.3. Heatmap generation and keypoint detection transformer

Transformer architectures have shown great ability in modeling discrete data as well as high-dimensional data such as images [4, 5]. The common architecture in these models usually comprise multi-head self-attention blocks which are followed by MLP feed-forward network (FFN) blocks. Our method here is inspired by [4, 31]. However, instead, here we investigate the representative power of quantized representations rather than CNN and Transformer descriptors. After training, VQ-VAE provides a codebook (C) that can be used for representing the frames of a sequence. These representations will be used as input to our temporal transformer architecture. First, the VQ-VAE backbone produces the frame level embeddings $f \in \mathcal{R}^{t \times d \times H \times W}$ ($t \in \{1, \ldots, T\}$ and d is embedding dimension). Because the Transformer encoder expects the inputs as a sequence, the embeddings are flattened to create a two-dimensional feature map. Due to the importance of order of the frames and to avoid permutation invariance problem of the Transformer architecture, the input features of each attention layer are complemented by positional encodings. The encoder follows a standard architecture consisting

of a multi-head self-attention module and a fully connected FFN that computes the encodings (E). The main purpose of the Transformer decoder is to decode pixel features that represent instances of objects in each frame. Like VisTR [31], we use a fixed number of input embeddings (instance queries) to query instance features, except that the number of our queries is much higher due to the larger number of instances per frame. Therefore, the decoder takes encoded VQ-VAE embedding from the encoder and n object instance queries and generates object instance level features. Similar to VisTR, the output of the decoder (instance level features) and encoded embeddings (E) are fed to an attention layer and concatenated with VQ-VAE embeddings. To produce the heatmaps, these features are then used to train a three layered 3D convolutional network followed by layer normalization and ReLU activation. In this way, the mask features of each object in different frames can get obtained. The network is optimized with a loss which is an extension of Focal loss [18] with a shape prior regularizer for encouraging the network to better learn the object shapes provided in the ground-truth:

$$\mathcal{L} = \alpha\mathcal{L}_f + \beta\mathcal{L}_{sh} \tag{5}$$

where α and β are hyperparameters determining weights of each loss in the optimization function. $\mathcal{L}_f$ is the extended Focal loss used for minimization of the estimates between the predicted heatmap $\hat{H}_{x,y}$ and its corresponding ground-truth heatmap $H_{x,y}$ and is defined as:

$$\mathcal{L}_f(H,\hat{H}) = -\frac{-1}{N}\sum_{xy}\begin{cases}(1-\hat{H}_{x,y})^{\gamma}\log(\hat{H}_{x,y}) & \text{if } H_{x,y}=1\\ (1-H_{x,y})^{\lambda}(\hat{H}_{x,y})^{\gamma} & \\ \log(1-\hat{H}_{x,y}) & \text{otherwise}\end{cases} \tag{6}$$

where γ and λ represent hyperparameters of the loss function and N is the total number of ground-truth object centers in the input image. $\mathcal{L}_f$ is our shape prior regularizer which helps in efficient generation of the heatmaps. By assuming that all the ground-truth objects are circular (Gaussians), the network is urged to adapt circular shapes in the heatmaps. Having c as the center of a circle-shaped object O, for any pixels p belonging to the object O, all the q pixels on the line l connecting p to the center c should be within the maximum range of r from c where r is the maximum diameter of the object annotations in the ground-truth. Therefore, the shape prior is defined as:

$$\mathcal{L}_{sh}(H,\hat{H}) = \sum_{p\in O}\sum_{q\in l} B_{p,q}\times D_{p,q}\times|P(\hat{H}_p|H) - P(\hat{H}_q|H)| \tag{7}$$

$B_{p,q}$ is 1 if both pixels are predicted as an object pixel and 0 if classified as background. $D_{p,q}$ is 1 while pixel q on line l is within range r from the center and 0 otherwise. $P(\hat{H}_p|H)$ and $P(\hat{H}_q|H)$ are the predicted probabilities of pixels p and q being an object in the predicted heatmap given the current heatmap ground-truth. Then, for inference, we take the local maximum values of the generated heatmaps as the center of the predicted keypoint coordinates of the available objects.

3.4. Imbalance problem and shape regularization in WAMI

Imbalance problem in object detection occurs when distribution of an input property affects the performance of the detector. In natural images, and in the most common form, the imbalance of the foreground-to-background class creates a severe disparity between the number of positive and negative samples. In this case, thousands of negative samples can be extracted against a few positive samples where ignoring this imbalance in the distribution of samples can greatly affect the detector and disrupt its performance. In WAMI, although there is a relatively higher number of objects per frame compared to natural images ($\sim 2k$), the imbalance problem is more pronounced as these objects occupy only less than 7×10^{-6} pixels of the total pixels in a frame [16]. Most regions are easy negatives without useful information that can overwhelm the training and there are only a few positive regions that are hard to differentiate (Fig. 1). While introducing a weight factor can balance the importance of positive versus negative samples, it has no effect on emphasizing differences between easy and hard examples (examples with large errors). Focal loss resolves this problem by reshaping the cross-entropy loss to down-weight easy samples and focus more on training hard negatives. This feature can also be valuable when it comes to WAMI images since the distribution of Gaussian peaks in heatmaps is unbalanced and covers a very small part of them. The scaling factor can be regularized by dynamically scaling the loss values. Therefore, by increasing confidence in the correct class, it can quickly decay to zero hence, automatically down-weight the effect of easy samples.

Furthermore, usually during training, optimization is performed only locally and in pixel-level of the input and the target, ignoring the global context and prior information. A shape prior, especially in the case of keypoint detection, provides global information on geometry of the target and guides the output mask to get closer to the shape of the class in the ground-truth. Given the similar shapes of the tiny targets occupying very small regions in the context of WAMI, this promotes the networks to learn easier and converge even faster.

4. Experiments

4.1. Dataset

In this work, we used the WPAFB 2009 dataset [1] to train and evaluate the proposed method. The images of

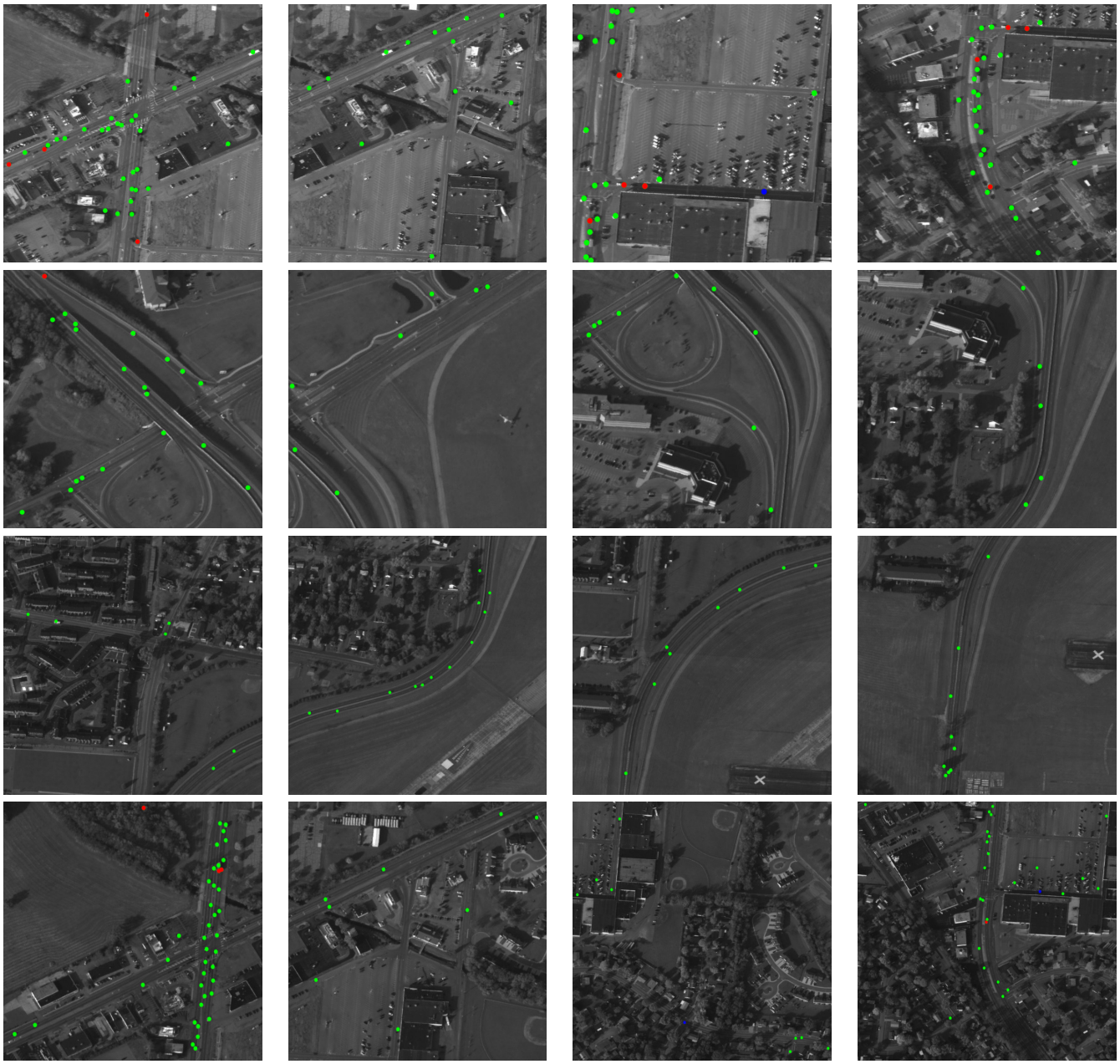

Figure 3. Visualization of qualitative detection results on different regions (Regions AOI 01, AOI 02, AOI 34, and AOI 40 from first to forth rows repectively). In the figures, True Positives (TP) are indicated in green, False Positives (FP) in red and False Negatives (FN) in blue.

the dataset were taken through six optical sensors and then the images were stitched together to cover a wide area of about 35 square kilometers. The dataset (1025 frames) is divided into two categories: training and testing, with 512 frames are dedicated for training and 513 frames for testing. Vehicles and their trajectory are manually annotated in each frame. Five different resolutions of the videos are included in the dataset, which we use the second largest image resolution ($\sim$ 13K $\times$11K pixels). Our framework is capable of detecting moving as well as stationary vehicles. In order to have a fair comparison with state-of-the-art, the evaluation protocols in [40] are followed.

4.2. Evaluations

All the proposal regions are divided to images of size 128×128 and normalized between $[-0.5, +0.5]$ prior to

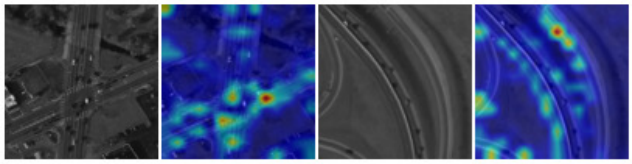

Figure 4. Illustrates attention maps of Transformer before 3DConv layers.

Metric	01	02	03	34	40	41
Precision	0.970	0.981	0.966	0.979	0.976	0.966
Recall	0.917	0.961	0.982	0.958	0.914	0.935
F1-score	0.943	0.970	0.973	0.968	0.944	0.950

Table 1. The precisions, recalls, and F_1 measures on six AOI using the proposed detection method.

the training. For training $60k$ ground-truth heatmaps are randomly generated ensuring that there is at least one object inside the boundary of the selected patch. The best performance achieved when the frame length is 5. All the models are developed in PyTorch [20] and optimized with Adam [15] with learning rate of 10^{-2} and 10^{-4} for proposal network and VQ-VAE with Transformer networks respectively. For VQ-VAE the Codebook size is 1024 with vector dimension of 256 and the dimension of latents are set to $8 \times 32 \times 32$. For an area of 128×128 the model can predicts up to 30 vehicles. Therefore, having a 5 frame sequence 150 object query is sent to the Transformer which comprises of 6 encoder and decoder layers with 4 attention heads. Transformer's hyperparameters are adjusted following VisTR. Batch size is set to 8 for all models where training and testing is done using two RTX 6000 GPUs with 192GB of RAM. A detected point is considered true positive (TP) when there is at least one ground-truth point in the 10-pixel range around that point. Detections outside of this range are considered as false positive (FP) and ground-truth annotation without any detection within its 10 pixels vicinity is considered as false negative (FN). Each annotated point in the ground-truth can only be assigned to one detection at inference. In case of multiple detections within the range only the closest one is considered as TP and the rest as FP (unless another annotation of the ground-truth is found that satisfies the conditions of a TP). The results are reported and compared in terms of precision, recall and F_1 measure. Table 1 shows the precision and recall of our proposed detector framework when applied to each AOI. For the most part, the detector achieves similar precision in different AOIs. However, for areas 03 and 41, the precision rate is a bit lower due to the detection of several non-vehicle moving objects which were not annotated in the ground-truth. Lower recall rates compared to precision is also the result of the incomplete ground-truth annotation

Method	**01**	**02**	**03**	**34**	**40**	**41**
[25]	0.645	0.760	0.861	–	–	–
[13]	0.743	0.825	0.876	0.763	0.737	0.708
[24]	0.783	0.793	0.876	0.755	0.749	0.762
[32]	0.738	0.820	0.868	0.761	0.733	0.700
[21]	0.816	0.868	0.892	–	–	–
[22]	0.850	0.876	0.889	0.826	0.817	0.799
[28]	0.866	0.890	0.900	–	–	–
[40]	0.935	0.947	0.945	0.953	0.935	0.934
[16]	**0.947**	0.951	0.942	0.933	**0.983**	0.928
[39]	0.944	0.967	0.964	0.967	0.938	0.948
Ours	0.943	**0.970**	**0.973**	**0.968**	0.944	**0.950**

Table 2. Comparison with state-of-the-art for six AOI in the WPAFB dataset in terms of F_1 measure. The top performance in each area is indicated in bold.

where only some stationary vehicles are arbitrarily annotated in the presence of many others in the surrounding. The visualization of the detection results are shown in figure 3, with each row containing examples of each region in the WPAFB dataset. It can be seen that the detection is performed well even in challenging situations where the object instances are dense or sparse. Figure 4 visualizes the output attention map of the Transformer which shows how it acquires more accurate neuron activations in the feature map. The maps may contain some local details which can cause noisy detections as some weak detection can get obscured by the neighbor activations. Retaining the local maximum value of the generated heatmaps and assigning zeros to the other positions prevents confusions during detection.

We compare our framework against state-of-the-art object detection methods in remote sensing which we are aware of. Table 2 shows the results in terms of detection accuracy using F_1 measure. The compared methods are from both background-subtraction-based [13, 24, 25, 40] and spatiotemporal-based [16, 39] categories where some use an extra post-processing step. Our proposed framework achieves competitive results compared to the others (highest accuracy in 4 out of 6 areas) given that it is anchor-free and does not use background subtraction or any post-processing technique.

4.3. Ablation study

This section provides extensive ablation experiments which are conducted to evaluate the key specifications of our framework.

Sequence length. The main difference between video and image stems from the temporal information contained in the videos. In spatiotemporal methods, the effective use of temporal information by the model is the most important factor in understanding videos. To show that, we train the model with different amount of temporal information. Table 3 demonstrates the effect of sequence lengths on the detector

performance and their comparison. Based on the results, as the length varies from 3 to 16, first there is an increase in the F_1 score and then, there is a monotone decrease. This is due to small input patch size and low frame rate of the videos. When the sequence length exceeds 5, almost none of the objects searched in the first frame are available, and stacking more frames does not help with better detection.

Codebook length. Table 4 shows experimental results of training the VQ-VAE codebook with different number of codes. It can be seen that increasing the number of codes improves the performance of the detector, however, further increase of the latent size does not affect the quality of detections in a positive way. This indicates that for this dataset, a medium size codebook is enough for representation of spatiotemporal information and a longer latent size surpasses the required base threshold for learning the representations.

Effect of feature encoding. To illustrate the effect of feature encoding, the detection module is trained by three types of input features: VQ-VAE encodings from the trained codebook, CNN-encoded, and Transformer-encoded features. As reported in Table 5 VQ-VAE encoded features achieve superior performance compared to both CNN-encoded and Transformer-encoded features in all AOIs. It is also interesting to see the superiority of the Transformer-encoded features to the CNN-encoded features. This might be related to the Transformer updates based on the pairwise similarities of its self-attention module which helps to learn spatial and temporal features together. The results show the power of discreet representations in modeling spatiotemporal information that achieve even better performance than the Transformer encoder which learns spatiotemporal information as a whole.

Loss function with shape prior. As discussed, imbalance problem poses a significant challenge for detector when the target object occupies considerably smaller amount of spatial space compared to the background. While cross-entropy based losses are being more affected by this problem, Focal loss solutions alleviate the issue by down-weighting the easy samples. In case of object detection formulated as a keypoint detection problem, adding a shape prior can be an effective solution. Table 6 compares the effect of different loss functions on the performance of the detection framework: cross-entropy loss, Focal loss, and Focal loss with a shape prior. Adding the shape prior significantly improves the performance and outperforms the detection in all regions when it is compared to the other loss function.

length	01	02	03	34	40	41
3	0.924	0.917	0.937	0.908	0.925	0.894
5	0.943	0.970	0.973	0.968	0.944	0.950
8	0.897	0.884	0.895	0.872	0.891	0.864
16	0.852	0.869	0.846	0.832	0.858	0.817

Table 3. **Sequence length.** Due to low frame rate, the performance (F_1 measure) drops as sequence length increases.

# of codes	01	02	03	34	40	41
256	0.864	0.886	0.904	0.891	0.874	0.919
1024	0.943	0.970	0.973	0.968	0.944	0.950
4096	0.938	0.953	0.959	0.927	0.946	0.922

Table 4. **Codebook length.** While increasing codebook size for embeddings improves the performance, further increase of code size does not help it.

Backbone	01	02	03	34	40	41
VQ-VAE	0.943	0.970	0.973	0.968	0.944	0.950
CNN	0.868	0.821	0.847	0.886	0.844	0.871
Transformer	0.907	0.885	0.894	0.865	0.883	0.898

Table 5. **VQ-VAE vs CNN vs. Transformer backbones.** VQ-VAE provides the best feature quality.

Loss	01	02	03	34	40	41
CE	0.883	0.913	0.904	0.892	0.914	0.921
FL	0.908	0.921	0.937	0.916	0.934	0.924
FL+SP	0.943	0.970	0.973	0.968	0.944	0.950

Table 6. **Loss function.** Cross-entropy vs Focal loss vs improved version of Focal loss with shape prior.

5. Conclusion

In this work, we proposed a novel framework for detecting tiny vehicles in WAMI sequences. We considered vehicles as keypoints in the relative background which are described by Gaussian distributions at the object position annotated in the ground-truth. To exploit the spatiotemporal information we trained a VQ-VAE network that learns a codebook of latent representations. To generate the detection masks, a Transformer network with a revamped loss function was trained which not only addresses the foreground-background imbalance problem but also tries to conform to the shape of the objects in the ground-truth benefiting from a shape regularizer. Extensive experiments demonstrated the performance of the proposed framework with ablation studies investigating different design choices. Compared to the existed methods that we are aware of, our detector achieves competitive results in the detection task. As the framework utilizes spatiotemporal information, we will explore joint detection and tracking in future research.

Acknowledgements This work has been supported by the BPI France funding under the LiChiE project. The authors are also grateful to the OPAL infrastructure provided by Université Côte d'Azur.

References

[1] US AFRL. Wright-patterson air force base (wpafb) dataset. https://www.sdms.afrl.af.mil/index.php?collection=wpafb2009, 2009. 5

[2] Wei Ao, Yanwei Fu, Xiyue Hou, and Feng Xu. Needles in a haystack: Tracking city-scale moving vehicles from continuously moving satellite. *IEEE Transactions on Image Processing*, 29:1944–1957, 2019. 1

[3] Songze Bao, Xing Zhong, Ruifei Zhu, Xiaonan Zhang, Zhuqiang Li, and Mengyang Li. Single shot anchor refinement network for oriented object detection in optical remote sensing imagery. *Ieee Access*, 7:87150–87161, 2019. 3

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 3, 4

[5] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, pages 1691–1703. PMLR, 2020. 4

[6] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. Memory enhanced global-local aggregation for video object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10337–10346, 2020. 3

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[8] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 4

[9] Liang Han, Pichao Wang, Zhaozheng Yin, Fan Wang, and Hao Li. Exploiting better feature aggregation for video object detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1469–1477, 2020. 3

[10] Fei He, Naiyu Gao, Qiaozhe Li, Senyao Du, Xin Zhao, and Kaiqi Huang. Temporal context enhanced feature aggregation for video object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10941–10948, 2020. 3

[11] Lu He, Qianyu Zhou, Xiangtai Li, Li Niu, Guangliang Cheng, Xiao Li, Wenxuan Liu, Yunhai Tong, Lizhuang Ma, and Liqing Zhang. End-to-end video object detection with spatial-temporal transformers. *arXiv preprint arXiv:2105.10920*, 2021. 3

[12] Licheng Jiao, Fan Zhang, Fang Liu, Shuyuan Yang, Lingling Li, Zhixi Feng, and Rong Qu. A survey of deep learning-based object detection. *IEEE access*, 7:128837–128868, 2019. 2

[13] Mark Keck, Luis Galup, and Chris Stauffer. Real-time tracking of low-resolution vehicles for wide-area persistent surveillance. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 441–448. IEEE, 2013. 1, 2, 7

[14] Phil Kent, Simon Maskell, Oliver Payne, Sean Richardson, and Larry Scarff. Robust background subtraction for automated detection and tracking of targets in wide area motion imagery. In *Optics and Photonics for Counterterrorism, Crime Fighting, and Defence VIII*, volume 8546, page 85460Q. International Society for Optics and Photonics, 2012. 2

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[16] Rodney LaLonde, Dong Zhang, and Mubarak Shah. Clusternet: Detecting small objects in large scenes by exploiting spatio-temporal information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4003–4012, 2018. 3, 4, 5, 7

[17] Pengpeng Liang, Haibin Ling, Erik Blasch, Guna Seetharaman, Dan Shen, and Genshe Chen. Vehicle detection in wide area aerial surveillance using temporal context. In *Proceedings of the 16th international conference on information fusion*, pages 181–188. IEEE, 2013. 1, 2

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3, 5

[19] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017. 4

[20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 7

[21] Thomas Pollard and Matthew Antone. Detecting and tracking all moving objects in wide-area aerial video. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 15–22. IEEE, 2012. 1, 7

[22] Vladimir Reilly, Haroon Idrees, and Mubarak Shah. Detection and tracking of large number of targets in wide area surveillance. In *European conference on computer vision*, pages 186–199. Springer, 2010. 1, 7

[23] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 4

[24] Imran Saleemi and Mubarak Shah. Multiframe many–many point correspondence for vehicle tracking in high density wide area aerial videos. *International journal of computer vision*, 104(2):198–219, 2013. 1, 2, 7

[25] Xinchu Shi, Haibin Ling, Erik Blasch, and Weiming Hu. Context-driven moving vehicle detection in wide area motion imagery. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 2512–2515. IEEE, 2012. 7

[26] Andrews Sobral and Antoine Vacavant. A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos. *Computer Vision and Image Understanding*, 122:4–21, 2014. 2

[27] Lars Sommer, Wolfgang Krüger, and Michael Teutsch. Appearance and motion based persistent multiple object tracking in wide area motion imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3878–3888, 2021. 4

[28] Lars Wilko Sommer, Michael Teutsch, Tobias Schuchert, and Jürgen Beyerer. A survey on moving object detection for wide area motion imagery. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–9. IEEE, 2016. 7

[29] Guanxiong Sun, Yang Hua, Guosheng Hu, and Neil Robertson. Mamba: Multi-level aggregation via memory bank for video object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2620–2627, 2021. 3

[30] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 3

[31] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8741–8750, 2021. 4, 5

[32] Jiangjian Xiao, Hui Cheng, Harpreet Sawhney, and Feng Han. Vehicle detection and tracking in wide field-of-view aerial video. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 679–684. IEEE, 2010. 1, 2, 7

[33] Chun-Han Yao, Chen Fang, Xiaohui Shen, Yangyue Wan, and Ming-Hsuan Yang. Video object detection via object-level temporal aggregation. In *European conference on computer vision*, pages 160–177. Springer, 2020. 3

[34] Yongtao Yu, Haiyan Guan, Dilong Li, Tiannan Gu, E Tang, and Aixia Li. Orientation guided anchoring for geospatial object detection from remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 160:67–82, 2020. 3

[35] Junpeng Zhang, Xiuping Jia, Jiankun Hu, and Kun Tan. Moving vehicle detection for remote sensing video surveillance with nonstationary satellite platform. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–1, 2021. 1

[36] Jiaxing Zhang, Chao Tao, and Zhengrong Zou. An on-road vehicle detection method for high-resolution aerial images based on local and global structure learning. *IEEE Geoscience and Remote Sensing Letters*, 14(8):1198–1202, 2017. 4

[37] Yuhang Zhang, Hao Sun, Jiawei Zuo, Hongqi Wang, Guangluan Xu, and Xian Sun. Aircraft type recognition in remote sensing images based on feature learning with conditional generative adversarial networks. *Remote Sensing*, 10(7):1123, 2018. 3

[38] Lin Zhou, Haoran Wei, Hao Li, Wenzhe Zhao, Yi Zhang, and Yue Zhang. Arbitrary-oriented object detection in remote sensing images based on polar coordinates. *IEEE Access*, 8:223373–223384, 2020. 3

[39] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European Conference on Computer Vision*, pages 474–490. Springer, 2020. 7

[40] Yifan Zhou and Simon Maskell. Detecting and tracking small moving objects in wide area motion imagery (wami) using convolutional neural networks (cnns). In *2019 22th International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE, 2019. 3, 6, 7

[41] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3