MULDE: Multiscale Log-Density Estimation via Denoising Score Matching for Video Anomaly Detection

Jakub Micorek Horst Possegger Dominik Narnhofer Horst Bischof Mateusz Koziński
Graz University of Technology, Austria
{jakub.micorek, possegger, dominik.narnhofer, bischof, mateusz.kozinski}@icg.tugraz.at

Abstract

We propose a novel approach to video anomaly detection: we treat feature vectors extracted from videos as realizations of a random variable with a fixed distribution and model this distribution with a neural network. This lets us estimate the likelihood of test videos and detect video anomalies by thresholding the likelihood estimates. We train our video anomaly detector using a modification of denoising score matching, a method that injects training data with noise to facilitate modeling its distribution. To eliminate hyperparameter selection, we model the distribution of noisy video features across a range of noise levels and introduce a regularizer that tends to align the models for different levels of noise. At test time, we combine anomaly indications at multiple noise scales with a Gaussian mixture model. Running our video anomaly detector induces minimal delays as inference requires merely extracting the features and forward-propagating them through a shallow neural network and a Gaussian mixture model. Our experiments on five popular video anomaly detection benchmarks demonstrate state-of-the-art performance, both in the object-centric and in the frame-centric setup.

1. Introduction

The goal of video anomaly detection (VAD) is to detect events that deviate from normal patterns in videos. VAD has numerous potential applications in healthcare, safety, and traffic monitoring. It can be used to detect events like human falling down, workplace, or traffic accidents, and holds the promise of dramatically reducing the time needed to respond to emergencies that can result from them. The main challenge of anomaly detection stems from the fact that, unlike classes of actions in video action recognition, anomalies do not form a coherent group of patterns and typically cannot be anticipated in advance. In consequence, in many applications anomalous training data is not available, necessitating the so-called one-class classification approach, in which the system is trained exclusively on normal data.

Traditional approaches to one-class video anomaly detection rely on training a deep network in auxiliary self-supervised tasks, like auto-encoding the frame sequence [9–12, 15, 29], predicting future frames [22, 28], inpainting spatio-temporal volumes [9], and solving jigsaw puzzles [2, 39]. The underlying assumption is that given a video sufficiently different from those of the training set, *i.e.* one containing an anomaly, the network should fail to complete the self-supervised task. However, the connection between data normality or abnormality and the performance of the network remains unclear. Deep networks can generalize beyond their training set, and there is no guarantee that anomalies make them fail to complete their task.

Our motivation is to lay more solid foundations for video anomaly detection. To that end, we treat feature vectors extracted from videos as realizations of a random variable with a fixed distribution, and seek to approximate its probability density function with a neural network. Such approximation would enable a principled and intuitive approach to detecting anomalies: since anomalous data is characterized by a low likelihood under the statistical model of normal data, it could be detected by thresholding the approximate density function.

Training a neural network to directly approximate $p(\mathbf{x})$, the probability density function of the training data is very challenging. However, Vincent [38] showed that injecting the data with zero-centered, iid Gaussian noise makes it easier to model the distribution of the noisy data $q(\tilde{\mathbf{x}})$. For sufficiently low levels of noise, q preserves the shape of p, which makes it a suitable basis for our anomaly indicator. Vincent's contribution consisted in proposing denoising score matching, a method to train a neural network to approximate $-\nabla_{\tilde{\mathbf{x}}} \log q(\tilde{\mathbf{x}})$, the negative log-gradient of the density function of noisy data, which became a core algorithm of a recent class of generative models [35]. We modify this method to train a neural anomaly indicator that approximates, up to a constant, the log-density, $-\log q(\tilde{\mathbf{x}})$, well suited to indicating anomalies thanks to its one-to-one relation to $q(\tilde{\mathbf{x}})$. Our approach is illustrated in Figure 1.

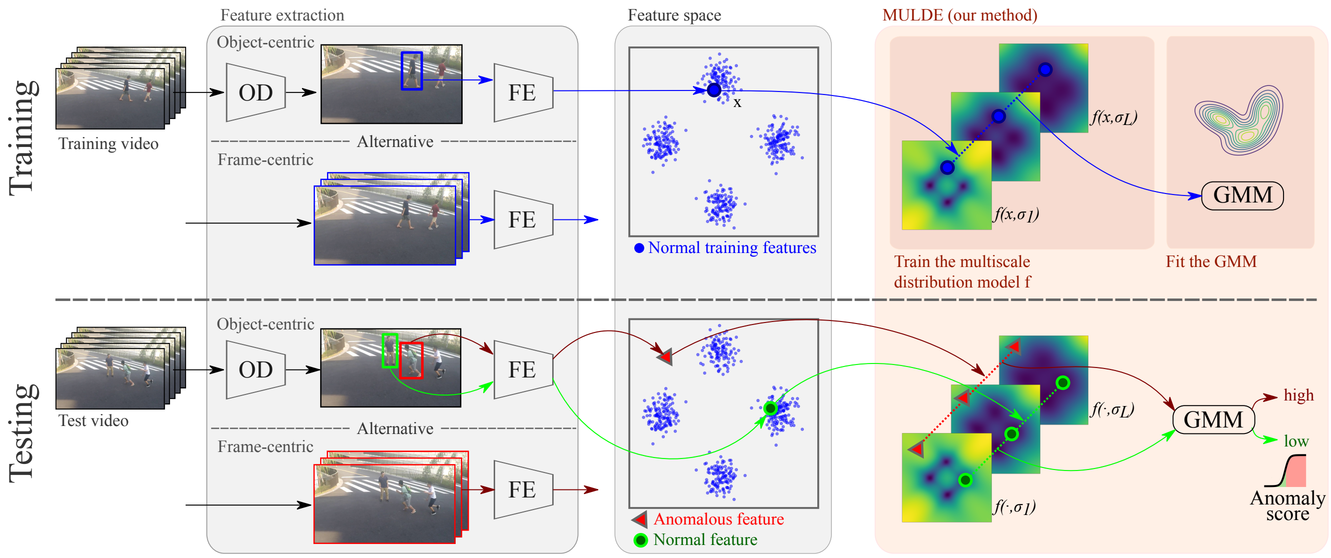

Figure 1. MULDE approximates the negative log-density of noisy, normal video features at multiple levels of noise σ with a neural network $f(\cdot, \sigma)$. The log-likelihoods estimated at multiple noise levels are combined into a single anomaly score with a Gaussian mixture model (GMM). MULDE can be trained to detect video anomalies in an object-centric or frame-centric manner. In the object-centric approach, an object detector (OD) is used to detect objects which are then fed to the feature extractor (FE). In the frame-centric approach, the feature extractor is applied to short sequences of entire frames.

In its basic form, introduced above, our method requires choosing the standard deviation σ of the noise injected into the data, also called the noise scale. This choice represents a compromise between making q closer to p for small values of σ and extending the support of q to cover more possible anomalies at larger noise levels. To avoid this unwelcome compromise, we do not settle on a single σ, but approximate the log-density for a range of noise scales $\sigma \in \{\sigma_1, \ldots \sigma_L\}$ with a neural network $f(\cdot, \sigma)$, and introduce a regularization term that tends to align the approximations at different scales. At test time, we compute anomaly indicators for a range of noise scales and combine them into a single anomaly score with a Gaussian mixture model, fitted to normal data. Our experiments show that **MULDE**, the regularized **MU**ltiscale **L**og-**DE**nsity approximation, is a very effective video anomaly detector.

To summarize, the main contribution of this paper is a novel approach to detecting anomalies from video features with a neural approximation of their log-density function. In technical terms, we propose a modification of multiscale denoising score matching for training anomaly indicators and a new method to regularize this training. Our anomaly detector is simple, mathematically sound, and fast at test time, as inference requires merely extracting the features and forward-propagating them through a neural network and a Gaussian mixture model. Moreover, it is agnostic of the feature vector it consumes on input. Our experiments on the Ped2 [26], Avenue [24], ShanghaiTech [25], UCFCrime [36], and UBnormal [1] data sets demonstrate state-of-the-art performance in anomaly detection both in the object-centric setup, where features are extracted from bounding boxes of detected objects, and in the frame-centric setup, with features computed for entire frames.

2. Related Work

VAD was studied in multiple settings: as a one-class classification problem, where no anomalous data is available for training [3–5, 7, 8, 12, 13, 24, 25, 28, 29, 32, 40, 42–45], as an unsupervised learning task, where anomalies are present in the training set, but it is not known which training videos contain them [45], and as a supervised, or weakly supervised problem, where training labels indicate anomalous video frames, or videos containing anomalies, respectively [1, 36, 45]. We address the first of these settings – we assume the training set is limited to normal videos.

Existing VAD methods can be categorized as frame-centric when they operate on features computed from entire frames or their sequences [1, 36, 40, 42, 45], and object-centric, if they estimate the abnormality of each bounding box in every frame [2, 8–10, 15, 32, 39], typically using a pre-trained feature extractor. The frame-centric design is more suited for global events, like fires, or smoke, while the object-centric one is oriented at anomalies associated with people or objects, like human falls, or vehicle accidents. In Sec. 4, we show that our method can establish state-of-the-art performance with features of either type.

The predominant approach to VAD is to train a deep network to auto-encode normal videos and use the reconstruction error as anomaly indicator [3, 7, 12, 25, 28, 29, 43, 44]. The idea of using the error of a model pre-trained on normal data to detect anomalies was extended from auto-encoding to multiple other tasks, including predicting future frames, or the optical flow [7, 21, 22, 28, 44], inpainting spatio-temporal volumes [9], and solving jigsaw puzzles [2, 39]. This overarching approach is predicated on the assumption that the error is higher for anomalous frames than for normal frames. However, there is no certainty that this assumption holds: it is not well understood under what conditions a neural network fails to perform its task and there is no guarantee that all anomalies make it fail. By contrast, no heuristic assumptions underlie the functioning of MULDE.

The idea of detecting video anomalies by modeling the distribution of normal video features recurs in the literature, but, to date, effective modeling techniques remain elusive. Some methods, like Gaussian mixture models [32], one-class Support Vector Machines [4, 5, 24], or multilinear classifiers [40], may lack the expressive power needed to reflect the complex and high-dimensional distribution of video features. Adversarially trained models [10, 20] offer high expressive power, but cannot guarantee to fully capture the distribution, because parts of the feature space may remain unexplored by the generator-discriminator pair during training. Diffusion models capture data distribution, but their use in VAD consists in generating samples of normal frames [42], or human poses [8], and comparing observed frames or poses to generated ones. This requires multiple diffusion steps and reduces the anomaly measure to a distance between the observation and the sample. Normalizing flows, recently used for detecting anomalies in human pose features [13], are free from these drawbacks as they explicitly approximate the likelihood of the training data. However, their performance decreases in the presence of complex correlations between features [18]. In contrast to these methods, MULDE combines all key ingredients of a VAD approach: high expressive power, the capacity to fully capture the distribution, and to accommodate arbitrary features.

The log-density approximation f, employed by MULDE to model the distribution of normal data, is often called the energy, in reference to the energy-based models [19], which represent probability distributions in the Boltzmann form $q(x) = \frac{1}{Z}e^{-f(x)}$, but restrict the model to the energy function f, since computing the normalization constant Z is typically infeasible. MULDE can therefore be seen as an energy-based model. For training f, MULDE relies on a modification of denoising score matching [38], a method to train a neural network to approximate the energy gradient. Score matching models the distribution of the training data injected with iid Gaussian noise, and Song and Ermon [35] extended it to multiple noise levels. Mahmood *et al.* [27] used the norm of this multi-level energy gradient approximation to detect anomalies in images. However, the gradient indicates all stationary points of the log-density function, which may appear both at the modes of the distribution, where normal data is concentrated, and in low-density regions, where anomalous data may reside. As shown in Fig. 2, some minima and maxima of the distribution remain indistinguishable even across a range of noise scales. By contrast, MULDE approximates the log-density function which, unlike its gradient, is a good anomaly indicator.

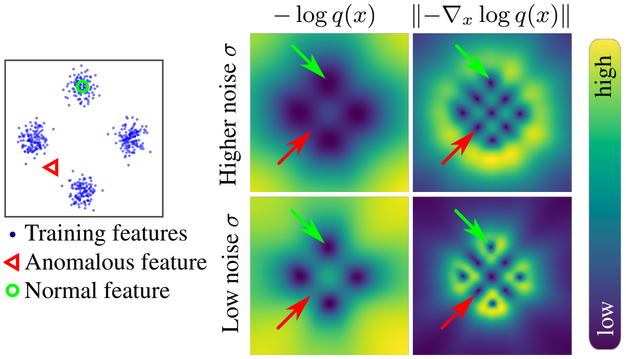

Figure 2. The log-density function is well suited for indicating anomalies, but its gradient is not. (*Left:*) A sample from a mixture of 4 Gaussians. (*Right:*) Learned negative log-density approximation (left column) and the norm of its gradient (right column). The negative log-density is a good anomaly indicator, taking low values for normal data and higher values for anomalous data. By contrast, the log-gradient norm is low not only at the modes of the distribution, but also at its minima between the modes, making it impossible to distinguish some anomalies from normal data.

3. Method

We perform anomaly detection in the space of semantic features extracted from videos. This lets us focus on detecting semantic anomalies, for example, unusual actions involving objects observed also under normal conditions, as opposed to anomalies in the space of raw input, like frame sequences that do not resemble real videos. We delegate feature extraction to off-the-shelf models and focus on the effective detection of anomalous events that they encode.

Motivation Intuitively, anomalous video features should not be observed under ‘normal’ conditions. More formally, an anomalous feature is characterized by a small likelihood under the assumed statistical model of anomaly-free video features. This suggests an approach to detecting anomalies by approximating the probability density function p and simply declaring test features with sufficiently low probability anomalous. We adopt this approach and model the probability density with a neural network.

Overview In practice, it is difficult to train a neural network to directly approximate the probability density of its training data p. However, injecting the data with noise makes this approximation feasible, as we will explain below. The distribution of the noisy data takes the form

$$q(\tilde{\mathbf{x}}) = \int \rho(\tilde{\mathbf{x}}|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}, \tag{1}$$

where $\rho(\tilde{\mathbf{x}}|\mathbf{x})$ denotes the conditional distribution of a noisy sample $\tilde{\mathbf{x}}$ given a noise-free sample $\mathbf{x}$, which we take to be an iid Gaussian centered at $\mathbf{x}$. The main idea behind our anomaly detector is that q preserves the shape of p but, in contrast to it, yields itself to a neural approximation. Specifically, we approximate, up to a constant, the negative log-density function $-\log q(\tilde{\mathbf{x}})$, which is an excellent anomaly indicator due to its bijective relation to $q(\tilde{\mathbf{x}})$. Low values of $-\log q(\tilde{\mathbf{x}})$ correspond to high probability density and are characteristic of normal data. Its high values indicate areas of low probability density where anomalous data may reside. Fig. 2 illustrates this on a synthetic example. The technique we use to approximate the log-density with a neural network is a modification of score matching [38], a method to approximate the log-density gradient.

We introduce score matching in Sec. 3.1 and our training method in Sec. 3.2. In Sec. 3.3 and 3.4, we extend this method to approximating the negative log-density at different scales of injected noise, and introduce a regularization term intended to facilitate combining the multi-scale approximations. We discuss the choice of video features in Sec. 3.5. A diagram of our approach is presented in Fig. 1.

3.1. Background: Denoising score matching

Vincent [38] proposed a method to train a neural network s, parameterized with a vector θ, to approximate the gradient of the negative log-density function of data perturbed with iid Gaussian noise, in the sense of solving

$$\min_{\theta} \mathbb{E}_{\tilde{\mathbf{x}} \sim q(\tilde{\mathbf{x}})} \left\| s_\theta(\tilde{\mathbf{x}}) + \nabla_{\tilde{\mathbf{x}}} \log q(\tilde{\mathbf{x}}) \right\|_2^2 . \tag{2}$$

This gradient approximation forms the foundation of a family of generative models [35] which initialize a sample with noise and use the log-gradient approximation to drive the sample close to the mode of the distribution. We will show that it also enables effective detection of video anomalies.

Directly evaluating the objective (2) is impossible because $q(\tilde{\mathbf{x}})$ is not known analytically, but Vincent [38] showed that it is equivalent, up to a constant, to

$$\min_{\theta} \mathbb{E}_{\substack{\mathbf{x} \sim p(\mathbf{x}) \\ \tilde{\mathbf{x}} \sim \mathcal{N}(\tilde{\mathbf{x}}|\mathbf{x}, \sigma \mathbf{I})}} \left\| s_\theta(\tilde{\mathbf{x}}) - \frac{\tilde{\mathbf{x}} - \mathbf{x}}{\sigma^2} \right\|_2^2, \tag{3}$$

which can be evaluated effectively. This gives rise to a stochastic algorithm for training s that iterates: composing a batch of noise-free training data $\mathbf{x}$, perturbing it with Gaussian noise to obtain a batch of noisy data $\tilde{\mathbf{x}}$, and making a gradient step on the expectation in Eq. (3), evaluated for the batch. Since this resembles training s to predict the noise injected to $\mathbf{x}$, it is often called *denoising* score matching.

3.2. Anomaly detection by denoising score matching

We modify the denoising score matching formulation to train a neural network f_θ, where θ denotes the vector of parameters, to approximate $-\log q(\tilde{\mathbf{x}})$, as opposed to its gradient. To that end, we change the objective (3), to train the *gradient* of f, instead of the network itself, which yields

$$\min_{\theta} \mathbb{E}_{\substack{\mathbf{x} \sim p(\mathbf{x}) \\ \tilde{\mathbf{x}} \sim \mathcal{N}(\tilde{\mathbf{x}}|\mathbf{x}, \sigma \mathbf{I})}} \left\| \nabla_{\tilde{\mathbf{x}}} f_\theta\left(\tilde{\mathbf{x}}\right) - \frac{\tilde{\mathbf{x}} - \mathbf{x}}{\sigma^2} \right\|_2^2 . \tag{4}$$

This makes $\nabla_{\tilde{\mathbf{x}}} f_\theta\left(\tilde{\mathbf{x}}\right)$ approximate $-\nabla_{\tilde{\mathbf{x}}} \log q(\tilde{\mathbf{x}})$ and, by the fundamental theorem of calculus, aligns f with $-\log q(\tilde{\mathbf{x}})$ up to a constant. Here, $f : \mathbb{R}^d \to \mathbb{R}$ is a mapping from the space of video features to scalar log-density values, and $\nabla_{\tilde{\mathbf{x}}} f_\theta : \mathbb{R}^d \to \mathbb{R}^d$, like s in the standard score matching formulation (3), maps d-dimensional video features to d-dimensional vectors of log-density gradients.

Notably, our formulation has an advantage over the standard denoising score matching even when the goal is to approximate the gradient of the log-density, as opposed to the log-density itself. By the Stokes' theorem, gradients form conservative – that is, curl-free – vector fields. Directly approximating the log-density gradient with a neural network, as done by the standard approach, may result in a vector field that is not conservative, in other words, does not represent a gradient of any function. By contrast, the gradient of a neural network trained using our formulation is guaranteed to form a conservative vector field. On the downside, training f_θ with our loss requires the network to be twice differentiable, which precludes the use of ReLU, several other nonlinearities, and max pooling.

3.3. Distribution modeling across noise scales

We recall that training with our loss (4) makes f_θ approximate the negative log-density of the distribution of noisy data $q(\tilde{\mathbf{x}})$, connected to the distribution of noise-free data $p(\mathbf{x})$ through the noise distribution $\rho(\tilde{\mathbf{x}}|\mathbf{x})$. ρ is an iid Gaussian centered at $\mathbf{x}$ and with a standard deviation σ. The choice of σ, called the noise scale, represents a compromise between making q closer to p at small noise scales and extending its support to cover more anomalies for larger noise levels. In theory, the optimal noise scale could be selected by cross-validation on a combination of normal and anomalous data, but in practice, anomalous validation data is rarely available. Therefore, instead of settling for a single noise scale, we approximate the log-density for a range of noise scales and combine the estimates at different scales by modeling their joint distribution with a Gaussian mixture.

To implement the multiscale log-density approximation, we take inspiration from Song and Ermon [35], who extended the original score matching formulation, presented in Eq. (3), to multiple noise scales, and apply a similar extension to our objective (4). Instead of approximating

$-\log q(\tilde{\mathbf{x}})$ for a fixed σ, we approximate a family of functions $-\log q_\sigma(\tilde{\mathbf{x}})$, parameterized by σ, with a neural network f_θ, conditioned on σ. We found it beneficial to put more emphasis on smaller values of σ when training f_θ. Thus, we sample σ from the log-uniform distribution on the interval $[\sigma_{\text{low}}, \sigma_{\text{high}}]$ and minimize

$$\min_{\theta} \mathbb{E}_{\substack{\mathbf{x}\sim p(\mathbf{x})\\ \tilde{\mathbf{x}}\sim\mathcal{N}(\tilde{\mathbf{x}}|\mathbf{x},\sigma\mathbf{I})\\ \sigma\sim\mathcal{LU}(\sigma_{\text{low}},\sigma_{\text{high}})}} \lambda(\sigma) \left\| \nabla_{\tilde{\mathbf{x}}} f_\theta\left(\tilde{\mathbf{x}}, \sigma\right) - \frac{\tilde{\mathbf{x}} - \mathbf{x}}{\sigma^2} \right\|_2^2, \qquad (5)$$

where $\lambda(\sigma)$ is a factor that balances the influence of the loss terms at different noise levels. We set $\lambda(\sigma) = \sigma^2$.

Once the network is trained, we fit a Gaussian mixture model to multi-scale log-density approximation vectors $[f_\theta(\mathbf{x}, \sigma_i)]_{i=1\dots L}$ for an evenly spaced sequence of noise levels $\{\sigma_i\}_{i=1}^{L}$, where $\sigma_1 = \sigma_{\text{low}}$ and $\sigma_L = \sigma_{\text{high}}$. At test time, our neural network takes a vector of video features and produces a multi-scale vector of log-density approximations, which is then input to the Gaussian mixture model yielding the final anomaly score.

3.4. Multiscale training regularization

The limitation of our method is that $f_\theta(\cdot, \sigma)$ can be trained to approximate $-\log q_\sigma(\tilde{\mathbf{x}})$ only up to a constant. That is, $f_\theta(\tilde{\mathbf{x}}, \sigma)$ effectively approximates $-\log q_\sigma(\tilde{\mathbf{x}}) + C_\sigma$, where C_σ is a constant that we do not know. In our formulation, there is no guarantee that this constant does not change across the range of σ. Since the variation of C_σ may make it more difficult to aggregate the estimates at different scales, we discourage it by using a regularization term $f_\theta(\mathbf{x}, \sigma)^2$, that penalizes the log-densities of noise-free examples. Our full training objective thus becomes

$$\min_{\theta} \mathbb{E}_{\substack{\mathbf{x}\sim p(\mathbf{x})\\ \tilde{\mathbf{x}}\sim\mathcal{N}(\tilde{\mathbf{x}}|\mathbf{x},\sigma\mathbf{I})\\ \sigma\sim\mathcal{LU}(\sigma_{\text{low}},\sigma_{\text{high}})}} \Bigg[\lambda(\sigma) \left\| \nabla_{\tilde{\mathbf{x}}} f_\theta\left(\tilde{\mathbf{x}}, \sigma\right) - \frac{\tilde{\mathbf{x}} - \mathbf{x}}{\sigma^2} \right\|_2^2 + \beta f_\theta(\mathbf{x}, \sigma)^2 \Bigg], \qquad (6)$$

where β is a hyperparameter of our method. Minimizing (6) no longer makes $\nabla_{\mathbf{x}} f_\theta$ an unbiased estimate of the log-gradient of the distribution, but as shown in our ablation studies, it improves our results in video anomaly detection. Algorithm 1 summarizes training the log-density approximation f_θ. The Gaussian mixture model is fitted with the standard expectation-maximization algorithm.

3.5. Selection of video features

Feature selection is closely tied to the type of target anomalies. For example, human pose features are well suited for detecting falls, and optical-flow-based ones help detect objects moving with unusual speeds or in unusual directions. MULDE is feature agnostic and in Sec. 4 we demonstrate its application with pose, velocity, and deep features, extracted from bounding boxes of object proposals, and ones extracted from entire frames and their sequences. As reported in Sec. 4, feature extraction dominates the running time of our method, which enables selecting the feature extractor to match the desired frame rate.

Algorithm 1 Training MULDE's anomaly indicator. The terms and steps that differ from the standard multi-scale denoising score matching [35] are highlighted.

```
Require:
    $f_\theta$ neural log-density model, parameterized by $\theta$
    $\mathcal{T}$ training set of normal video features
    $\sigma_{\text{low}}$ and $\sigma_{\text{high}}$, limits of the noise scale range
    $\beta$ regularization strength
$\theta \leftarrow$ random initialization
while not converged do
    $\mathcal{X} \leftarrow$ sample a batch from $\mathcal{T}$
    for $\mathbf{x} \in \mathcal{X}$ do                          ▷ for each batch element
        $\sigma \leftarrow$ sample log-uniform$(\sigma_{\text{low}}, \sigma_{\text{high}})$
        $\tilde{\mathbf{x}} \leftarrow$ sample $\mathcal{N}(\mathbf{x}, \sigma\mathbf{I})$
        Compute $f_\theta\,(\tilde{\mathbf{x}}, \sigma)$ by forward-propagation
        Compute $f_\theta\,(\mathbf{x}, \sigma)$ by forward-propagation
        Compute $\nabla_{\tilde{\mathbf{x}}} f_\theta\,(\tilde{\mathbf{x}}, \sigma)$ by backpropagation
        $\mathcal{L}(\mathbf{x}) \leftarrow \sigma^2 \left\| \nabla_{\tilde{\mathbf{x}}} f_\theta\,(\tilde{\mathbf{x}}, \sigma) - \frac{\tilde{\mathbf{x}}-\mathbf{x}}{\sigma^2} \right\|_2^2 + \beta f_\theta\,(\mathbf{x}, \sigma)^2$
    end for
    $\mathcal{L} \leftarrow \frac{1}{|\mathcal{X}|} \sum_{\mathbf{x}\in\mathcal{X}} \mathcal{L}(\mathbf{x})$          ▷ aggregate loss over batch
    Compute $\nabla_\theta \mathcal{L}$ by backpropagation
    Update $\theta$ using $\nabla_\theta \mathcal{L}$ with Adam
end while
```

4. Experimental Evaluation

Data sets We evaluated MULDE on five VAD benchmarks, containing videos captured with static cameras:

- **Ped2** [26] includes 16 training and 12 test videos of a campus scene. The training videos show pedestrians and the test videos contain anomalies, like cyclists, skateboarders, and cars in pedestrian areas.
- **Avenue** [24] consists of 16 training and 21 test videos of a walkway. Anomalies include people running, throwing objects, and walking in the wrong direction.
- **ShanghaiTech** [25] comprises 330 training and 107 test videos of 13 pedestrian traffic scenes differing by the camera viewpoint and lighting conditions. Anomalous events include robbery, jumping, fighting, and cycling.
- **UBnormal** [1] is composed of 543 synthetic videos of 29 virtual scenes. It contains human-related anomalies, like fighting, running, and jumping, but it also includes car accidents and environmental anomalies, like fog.
- **UCFCrime** [36] contains 1900 real-world surveillance videos, totaling 128 hours, and including 13 anomaly types, like fighting, robbery, road accidents, and burglary.

All our experiments were run in the ‘one-class classification’ setting, that is, we used no anomalous videos for training. In the experiments on **UBnormal** and **UCFCrime**, which contain anomalous training videos, we discarded these videos and restricted training to normal data.

Performance metrics We followed the standard practice and gauged performance on a frame-by-frame basis. For the object-centric approaches, which yield an anomaly score for each object detected in every frame, we took the highest, *i.e.* most anomalous, score in a frame for the evaluation. We used the area under the receiver operating characteristic curve (AUC-ROC) as the main performance metric.

Two methods to aggregate the AUC-ROC over multiple videos can be found in the literature: the *micro* and the *macro* score [1, 2, 10, 32, 33]. For the *macro* score, the AUC-ROC is computed separately for each video and then averaged across all videos in the test set. The *micro* score computes the AUC-ROC jointly for all frames of all test videos. In abstract terms, the *macro* score reflects performance attained by adjusting the detection threshold for each video independently, and the *micro* score is more conservative and applies the same threshold to all test videos. Since we address a VAD use case without an adaptive threshold, we rely on the *micro* score in our evaluation, but report the results in terms of both metrics. In the supplementary material, we additionally report the tracking- and region-based metrics by Ramachandra and Jones [31].

Baselines We compared MULDE to the best-performing VAD algorithms, including ones based on the reconstruction error [9, 15, 44], auxiliary tasks [1, 2, 9, 23, 33, 39, 42], adversarial training [1, 7, 10, 45], normalizing flows [13], and one-class classification with a multilinear classifier [40]. We reproduced the performance metrics of these methods as reported in the original papers. We present an even broader comparison, including less recent work, in the supplementary material.

Two baselines are related to our method more closely than the others. The AccI-VAD [32] approximates the probability density function of normal video features with a Gaussian mixture model, while we perform this approximation with a neural network. MSMA [27] uses the norm of the log-density gradient as an anomaly indicator, while our anomaly indicator is based on the log-density itself. Since MSMA was developed for image anomaly detection, we re-implemented it to work with video features.

Implementation details In our experiments, our density model f_θ parametrized by θ has two hidden layers with 4096 units followed by GELU nonlinearities. The final layer has an output dimension of one without any nonlinearity. It is trained using the Adam update rule [17], with exponential decay rates $\beta_1 = 0.5$ and $\beta_2 = 0.9$, and a batch size of 2048. We use the learning rates of 5e-4 and 1e-4 in the object- and frame-centric experiments, respectively.

Method	Ped2		Avenue		ShanghaiTech	
	Micro	Macro	Micro	Macro	Micro	Macro
CAE-SVM [15]	94.3	97.8	87.4	90.4	78.7	84.9
VEC [44]	97.3	-	90.2	-	74.8	-
SSMTL [9]	97.5	$\underline{99.8}$	91.5	91.9	82.4	89.3
HF^2 [23]	99.3	-	91.1	93.5	76.2	-
BA-AED [10]	98.7	99.7	92.3	90.4	82.7	89.3
[10]+SSPCAB [33]	-	-	92.9	91.9	83.6	89.5
Jigsaw-Puzzle [39]	99.0	**99.9**	92.2	93.0	84.3	89.8
[9]+UbNormal [1]	-	-	93.0	93.2	83.7	90.5
AccI-VAD [32]	99.1	**99.9**	93.3	**96.2**	85.9	89.6
SSMTL++ [2]	-	-	$\underline{93.7}$	92.5	83.8	90.5
MSMA* [27]	$\underline{99.5}$	**99.9**	90.2	92.5	84.1	90.2
STG-NF [13]	-	-	-	-	85.9	-
MULDE$_{\beta=0}$ (ours)	**99.7**	**99.9**	93.1	$\underline{96.1}$	$\underline{86.4}$	$\underline{91.0}$
MULDE (ours)	**99.7**	**99.9**	**94.3**	$\underline{96.1}$	**86.7**	**91.5**

Table 1. Object-centric results. Frame-level AUC-ROC (%) comparison (best marked **bold**, second best underlined). *implemented by us.

Both during training and at test time, we standardize the video features component-wise using the statistics of the training set. During training, we sample the noise scale used for each batch element from the log-uniform distribution on the interval $[\sigma_{\text{low}}, \sigma_{\text{high}}] = [\text{1e-3}, 1.0]$. For evaluation, we use $L = 16$ evenly spaced noise levels between $[\sigma_{\text{low}}, \sigma_{\text{high}}]$.

Video feature extraction MULDE is agnostic of the features used to represent the video content. In our experiments, we reused off-the-shelf video feature extractors.

In the object-centric experiments, we used the feature extraction pipeline proposed by Reiss and Hoshen [32]. It detects objects in each video frame and extracts deep and velocity features for each detected object. Additionally, pose features are extracted for each detected person. The pose vector contains coordinates of 17 keypoints and is obtained with AlphaPose [6]. CLIP image encoder [30] is used to extract deep features, which take the form of 512-dimensional vectors. The velocity features are produced by FlowNet 2.0 [14] and binned into histograms of oriented flows.

In the frame-centric experiments, we extracted features using Hiera-L [34], a masked autoencoder pre-trained on images and fine-tuned on Kinetics400 [16], a large-scale video action recognition data set. Hiera-L takes sequences of 16 frames as input and produces feature 1152-dimensional feature vectors.

Performance in object-centric VAD We present the results in object-centric VAD in Table 1. MULDE outperforms all the baselines in terms of the more conservative *micro* score. Interestingly, AccI-VAD which, like MULDE, relies on modeling the probability density of normal data,

Method	ShanghaiTech		UCF-Crime		UBnormal	
	Micro	Macro	Micro	Macro	Micro	Macro
CT-D2GAN [7]	77.7	-	-	-	-	-
CAC [41]	<u>79.3</u>	-	-	-	-	-
Scene-Aware [37]	74.7	-	72.7	-	-	-
GODS [40]	-	-	70.5	-	-	-
GCL [45]	-	-	74.2	-	-	-
UBnormal [1]	-	-	-	-	68.5	80.3
FPDM [42]	78.6	-	74.7	-	62.7	-
AccI-VAD$_{\text{GMM}}$* [32]	76.2	82.9	60.3	84.5	66.8	83.2
AccI-VAD$_{\text{kNN}}$* [32]	71.9	83.1	53.0	82.7	65.2	82.5
MSMA* [27]	76.7	84.2	64.5	83.4	70.3	85.1
MULDE$_{\beta=0}$(ours)	78.4	**86.0**	<u>75.9</u>	<u>84.8</u>	<u>71.3</u>	**86.0**
MULDE(ours)	**81.3**	<u>85.9</u>	**78.5**	**84.9**	**72.8**	<u>85.5</u>

Table 2. Frame-centric results. Frame-level AUC-ROC (%) comparison (best marked **bold**, second best <u>underlined</u>). *implemented by us.

β	ShanghaiTech		UCF-Crime		UBnormal	
	Micro	Macro	Micro	Macro	Micro	Macro
0.0	78.4	**86.0**	75.9	84.8	71.3	**86.0**
0.01	80.7	84.9	76.6	<u>84.9</u>	**72.9**	85.2
0.1	<u>81.3</u>	<u>85.9</u>	**78.5**	<u>84.9</u>	<u>72.8</u>	<u>85.5</u>
1.0	**81.4**	84.5	<u>77.2</u>	**85.5**	72.5	84.7

Table 3. Frame-centric performance of MULDE trained with different values of the regularization parameter β. Frame-level AUC-ROC (%) comparison (best marked **bold**, second best <u>underlined</u>).

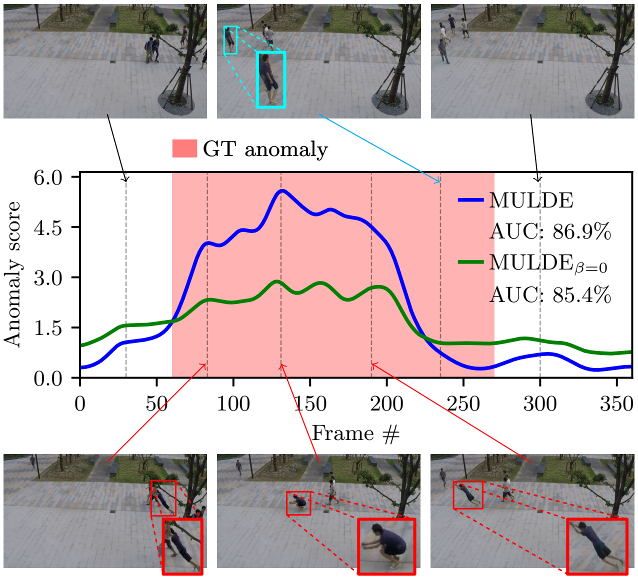

Figure 3. Anomaly detection with MULDE in a test video of the **ShanghaiTech** data set (video 13 in scene 4). Pedestrians walking in frames 30 and 300 represent normal behavior. A person jumping across the scene is annotated as anomalous. The anomaly indication produced by MULDE is aligned with the ground truth (GT) at its beginning but terminates earlier than the GT annotation. However, careful examination of the video reveals that normal behavior (walking, cyan bounding box in the top row) is re-instantiated before the end of the annotation, as indicated by MULDE. A regularized model produces a stronger anomaly indication (plotted in blue) than one without regularization (green plot).

ranks third on all three data sets. The high accuracy of both methods speaks in favor of the density modeling approach, while the edge MULDE holds over AccI-VAD attests to the superiority of MULDE's neural density model over the combination of Gaussian mixture models and the k-th nearest neighbor technique employed by AccI-VAD. MSMA, which uses an approximation of the log-density gradient as anomaly indicator, is outperformed by our log-density-based method by a fair margin.

Performance in frame-centric VAD As can be seen in Table 2, in frame-centric VAD, MULDE outperforms other methods on all three data sets. Notably, on the **UCFCrime** data set, by far the largest publicly available VAD benchmark, we advance the state of the art in terms of the *micro* score by 3.8 percent points. We improve the state of the art by more than 2pp on **ShanghaiTech** and **UBnormal**.

Moreover, while recent object-centric methods dominate frame-centric methods on data sets enabling both types of evaluation, MULDE narrows this performance gap, attaining the *micro* score of 81.3% in frame-centric VAD on **ShanghaiTech** and 86.7% in object-centric VAD on the same data set. On **UBnormal**, our frame-centric approach even outperforms the object-centric state-of-the-art method [13] by 1.1 percent points of the *micro* score.

Ablation studies To validate the design of MULDE, we run ablation studies and evaluated the contribution to performance from our regularization term, the Gaussian mixture model, and the architecture of our anomaly indicator.

As shown in Table 3, regularization improves the results of frame-centric VAD in terms of the *micro* score, although different values of the regularization factor β are optimal for different data sets. We take $\beta = 0.1$ as a compromise among performance on the three data sets. A qualitative example of anomaly detection by MULDE with and without regularization, shown in Fig. 3, demonstrates that a regularized model produces stronger anomaly indications.

As explained in Section 3.3, MULDE combines anomaly indications computed for a range of noise scales with a Gaussian mixture model fitted to normal data. This lets us avoid the need to select the noise scale to be used at test time. Fig. 4 demonstrates the impact of the mixture model on performance. It shows the *micro* score attained by our anomaly indicator $f_\theta(\cdot, \sigma)$ at individual noise scales, without the mixture model. The best performance is attained for intermediate levels of noise, which suggests a tradeoff between representing the noise-free distribution more faith-

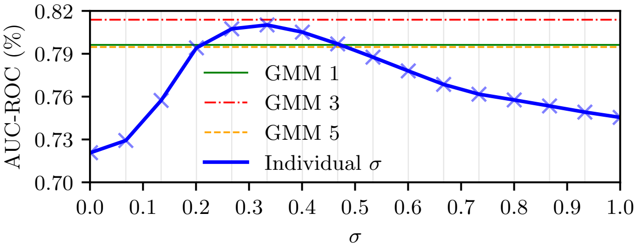

Figure 4. Performance of MULDE in frame-centric VAD on the **ShanghaiTech** data set without the Gaussian mixture model, with our anomaly indicator computed at individual noise scales (blue plot), and with the mixture model with 1, 3, and 5 components.

Units	1 Layer		2 Layers		3 Layers		4 Layers	
	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
1024	77.3	84.2	79.9	85.4	<u>81.2</u>	<u>85.8</u>	79.9	84.2
2048	77.6	84.6	80.4	84.6	79.2	83.5	79.4	84.0
4096	78.4	84.6	**81.3**	**85.9**	79.5	84.4	80.0	84.2
8192	79.2	84.3	79.4	84.8	80.4	84.4	79.3	84.7

Table 4. Performance of MULDE with architectural variations in frame-centric VAD on **ShanghaiTech**. Units indicate the number of neurons within a hidden layer. Columns correspond to the number of hidden layers. Frame-level AUC-ROC (%) comparison (best marked **bold**, second best <u>underlined</u>).

fully at lower noise scales and better covering the space away from training examples at higher noise scales. Performance attained with the mixture model slightly exceeds the one attained with the optimal noise scale, validating the GMM as a method to avoid tuning the noise level.

To find the optimal architecture of our anomaly indicator f_θ, we performed an ablation study over the number of hidden layers and the number of neurons in each hidden layer. The results, presented in Table 4, suggest the optimal depth of two hidden layers and the width of 4096 neurons, which we used in our experiments.

Running time Once a feature has been extracted from the video, MULDE requires only a forward-pass through the shallow network f_θ and an evaluation of a Gaussian mixture model. This represents a very small computational overhead over feature extraction. In the frame-centric approach with 512-dimensional feature vectors, MULDE takes less than one millisecond to process a single feature. For comparison, the Hiera-L feature extractor [34], used in the frame-centric experiments, requires 130 milliseconds to compute one feature vector from a sequence of 16 frames. MULDE's running time is therefore dominated by the time needed for feature extraction, which means that the feature extractor can be selected to match the target framerate on a given architecture. For example, in our setup, extracting a single feature with the smaller Hiera-base model takes 33 milliseconds, which enables video anomaly detection at 25 FPS, even on our PC with an NVIDIA RTX 2080Ti GPU.

5. Discussion

Limitations Our method has two main limitations. First, our video anomaly detector requires features of anomalous and normal videos to be distinguishable. Intuitively, this condition should be satisfied when the video feature extractor is selected adequately, but there is no theoretical guarantee that this is indeed the case. Second, even though our method eliminates the need to select the noise scale used for training, the range of employed noise scales should be sufficiently large to cover anomalies that would be seen at test time. Currently, there is no automatic way to select the upper limit of the noise range. We circumvent this limitation by standardizing the features component-wise and using a fixed, wide range of noise scales.

Future work We plan to extend this work by exploring architectures of the video anomaly indicator. In particular, reformulating the anomaly indicator in terms of the reconstruction error [9–12, 15, 29] represents a distinct opportunity to integrate previous methods into our approach.

Our method uses Gaussian noise during training, but other forms of noise may represent anomalies more effectively. For example, emulating random motion patterns in the video might be useful for detecting objects moving along unusual trajectories or with abnormal velocities.

Finally, we plan to extend MULDE from the one-class-classification scenario to the (weakly-) supervised one.

Conclusion We presented a novel approach to video anomaly detection by modeling the distribution of non-anomalous video features with a neural network. Our method is fast and attains state-of-the-art performance, both in the frame-centric and object-centric VAD. Most importantly, it is firmly grounded in statistical modeling. We defined abnormality in terms of the probability density under the distribution of normal data and designed MULDE to approximate the log-density function. This lets us discuss our method in statistical terms. For example, when our anomaly indicator fails to classify a video as anomalous, we know that its log-density approximation is not good enough. We can then plan actions to rectify this, for example, by changing the neural network architecture, acquiring more training data, or modifying the range of noise. By contrast, it might not be obvious what actions should be taken when previous methods fail, for example, by showing a small reconstruction error for an anomalous item.

Acknowledgements This work was partially funded by the Austrian Research Promotion Agency (FFG) under the project High-Scene (884306) and by the Austrian Science Fund (FWF) Lise Meitner grant (M3374).

References

[1] Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnormal: New benchmark for supervised open-set video anomaly detection. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[2] Antonio Barbalau, Radu Tudor Ionescu, Mariana-Iuliana Georgescu, Jacob Dueholm, Bharathkumar Ramachandra, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B. Moeslund, and Mubarak Shah. Ssmtl++: Revisiting self-supervised multi-task learning for video anomaly detection. *Computer Vision and Image Understanding*, 229, 2023.

[3] Yunpeng Chang, Zhigang Tu, Wei Xie, and Junsong Yuan. Clustering driven deep autoencoder for video anomaly detection. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2020.

[4] Rizwan Chaudhry, Avinash Ravichandran, Gregory Hager, and René Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[5] Rensso Victor Hugo Mora Colque, Carlos Caetano, Matheus Toledo Lustosa de Andrade, and William Robson Schwartz. Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos. *Transactions on Circuits and Systems for Video Technology*, 27(3):673–682, 2016.

[6] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proc. of the International Conference on Computer Vision (ICCV)*, 2017.

[7] Xinyang Feng, Dongjin Song, Yuncong Chen, Zhengzhang Chen, Jingchao Ni, and Haifeng Chen. Convolutional transformer based dual discriminator generative adversarial networks for video anomaly detection. In *Proc. of the International Conference on Multimedia*, 2021.

[8] Alessandro Flaborea, Luca Collorone, Guido Maria D'Amely di Melendugno, Stefano D'Arrigo, Bardh Prenkaj, and Fabio Galasso. Multimodal motion conditioned diffusion model for skeleton-based video anomaly detection. In *Proc. of the International Conference on Computer Vision (ICCV)*, 2023.

[9] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[10] Mariana Iuliana Georgescu, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. A background-agnostic framework with adversarial training for abnormal event detection in video. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 44(9), 2021.

[11] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proc. of the International Conference on Computer Vision (ICCV)*, 2019.

[12] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[13] Or Hirschorn and Shai Avidan. Normalizing flows for human pose anomaly detection. In *Proc. of the International Conference on Computer Vision (ICCV)*, 2023.

[14] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[15] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[16] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2015.

[18] Polina Kirichenko, Pavel Izmailov, and Andrew G Wilson. Why normalizing flows fail to detect out-of-distribution data. In *Proc. of the Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[19] Yann LeCun, Sumit Chopra, Raia Hadsell, Marc Aurelio Ranzato, and Fu jie Huang. A tutorial on energy-based learning. *Predicting Structured Data*, 1, 2006.

[20] Sangmin Lee, Hak Gu Kim, and Yong Man Ro. Stan: Spatio- temporal adversarial networks for abnormal event detection. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[21] Sangmin Lee, Hak Gu Kim, and Yong Man Ro. Bman: Bidirectional multi-scale aggregation networks for abnormal event detection. *Transactions on Image Processing*, 29, 2019.

[22] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[23] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *Proc. of the International Conference on Computer Vision (ICCV)*, 2021.

[24] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proc. of the International Conference on Computer Vision (ICCV)*, 2013.

[25] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proc. of the International Conference on Computer Vision (ICCV)*, 2017.

[26] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[27] Ahsan Mahmood, Junier Oliva, and Martin Andreas Styner. Multiscale score matching for out-of-distribution detection. *Proc. of the International Conference on Learning Representations (ICLR)*, 2021.

[28] Trong-Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *Proc. of the International Conference on Computer Vision (ICCV)*, 2019.

[29] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. of the International Conference on Machine Learning (ICML)*, 2021.

[31] Bharathkumar Ramachandra and Michael Jones. Street scene: A new dataset and evaluation protocol for video anomaly detection. In *Proc. of the Winter Conference on Applications of Computer Vision (WACV)*, 2020.

[32] Tal Reiss and Yedid Hoshen. Attribute-based representations for accurate and interpretable video anomaly detection. *arXiv preprint arXiv:2212.00789*, 2022.

[33] Nicolae-Cătălin Ristea, Neelu Madan, Radu Tudor Ionescu, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B Moeslund, and Mubarak Shah. Self-supervised predictive convolutional attentive block for anomaly detection. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[34] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li, and Christoph Feichtenhofer. Hiera: A hierarchical vision transformer without the bells-and-whistles. *Proc. of the International Conference on Machine Learning (ICML)*, 2023.

[35] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Proc. of the Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[36] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[37] Che Sun, Yunde Jia, Yao Hu, and Yuwei Wu. Scene-aware context reasoning for unsupervised abnormal event detection in videos. In *Proc. of the International Conference on Multimedia*, 2020.

[38] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

[39] Guodong Wang, Yunhong Wang, Jie Qin, Dongming Zhang, Xiuguo Bao, and Di Huang. Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2022.

[40] Jue Wang and Anoop Cherian. Gods: Generalized one-class discriminative subspaces for anomaly detection. In *Proc. of the International Conference on Computer Vision (ICCV)*, 2019.

[41] Ziming Wang, Yuexian Zou, and Zeming Zhang. Cluster attention contrast for video anomaly detection. In *Proc. of the International Conference on Multimedia*, 2020.

[42] Cheng Yan, Shiyu Zhang, Yang Liu, Guansong Pang, and Wenjun Wang. Feature prediction diffusion model for video anomaly detection. In *Proc. of the International Conference on Computer Vision (ICCV)*, 2023.

[43] Zhiwei Yang, Peng Wu, Jing Liu, and Xiaotao Liu. Dynamic local aggregation network with adaptive clusterer for anomaly detection. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2022.

[44] Guang Yu, Siqi Wang, Zhiping Cai, En Zhu, Chuanfu Xu, Jianping Yin, and Marius Kloft. Cloze test helps: Effective video anomaly detection via learning to complete video events. In *Proc. of the International Conference on Multimedia*, 2020.

[45] M Zaigham Zaheer, Arif Mahmood, M Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee. Generative cooperative learning for unsupervised video anomaly detection. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.