Video Anomaly Detection via Sequentially Learning Multiple Pretext Tasks

Chenrui Shi[1,2], Che Sun[2,1]*, Yuwei Wu[1], Yunde Jia[2]
[1]Beijing Key Laboratory of Intelligent Information Technology,
School of Computer Science & Technology, Beijing Institute of Technology, China
[2]Guangdong Laboratory of Machine Perception and Intelligent Computing,
Shenzhen MSU-BIT University, China
{shichenrui,sunche,wuyuwei,jiayunde}@bit.edu.cn

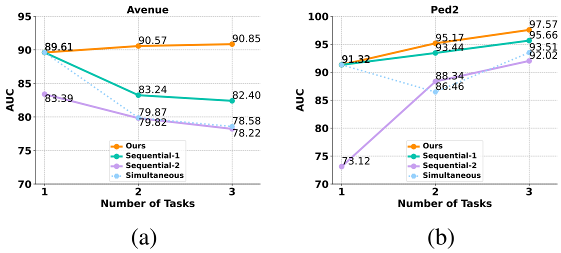
Abstract

Learning multiple pretext tasks is a popular approach to tackle the nonalignment problem in unsupervised video anomaly detection. However, the conventional learning method of simultaneously learning multiple pretext tasks, is prone to sub-optimal solutions, incurring sharp performance drops. In this paper, we propose to sequentially learn multiple pretext tasks according to their difficulties in an ascending manner to improve the performance of anomaly detection. The core idea is to relax the learning objective by starting with easy pretext tasks in the early stage and gradually refine it by involving more challenging pretext tasks later on. In this way, our method is able to reduce the difficulties of learning and avoid converging to sub-optimal solutions. Specifically, we design a tailored sequential learning order for three widely-used pretext tasks. It starts with frame prediction task, then moves on to frame reconstruction task and last ends with frame-order classification task. We further introduce a new contrastive loss which makes the learned representations of normality more discriminative by pushing normal and pseudo-abnormal samples apart. Extensive experiments on three datasets demonstrate the effectiveness of our method.

1. Introduction

Existing unsupervised video anomaly detection methods train models to perform a single pretext task such as frame reconstruction [14] or frame prediction [21], and they can discriminate anomalies when videos are significantly deviant from model expectations. These methods often render sub-optimal performances due to the nonalignment [17] between the single pretext task and video anomaly detection.

Recent methods [9, 17] resort to multiple pretext tasks to tackle the nonalignment problem, as multiple pretext tasks can provide more comprehensive and informative guidance than one single pretext task. The learning method of multiple pretext tasks is significant, yet under-explored. Conventional learning method, *i.e.*, simultaneously learning multiple pretext tasks, could not bring about the expected performance gains and even cause sharp performance drops. An example is shown in Fig. 1. The performance curves (the dotted lines in blue) plummet or fluctuate with more pretext tasks, and neither of the performance curves reaches the summit when leveraging all pretext tasks. The main reason is that models tend to get stuck in pareto-optimal points when simultaneously learning multiple pretext tasks. The pareto-optimal points [7, 34, 12] are such points that we could not further optimize any of the pretext task objectives without compromising the rest, striking a trade-



(a) (b)

Figure 1. AUC (%) performances of anomaly detection when learning multiple pretext tasks simultaneously and sequentially. The dotted line denotes simultaneously learning multiple pretext tasks. The solid lines denote sequentially learning multiple pretext tasks. The learning order of our method is “prediction (Pre) - reconstruction (Rec) - classification (Cls)”. The learning orders of Sequential-1 and Sequential-2 are “Pre-Rec-Cls” and “Rec-Cls-Pre” respectively. AUC performances on Avenue [24] and Ped2 [28] datasets show that the learning methods and the learning orders of multiple pretext tasks significantly influences the trained model’s ability for video anomaly detection.

[*]Corresponding author: Che Sun

off among them. Pareto-optimal points are not necessarily good solutions for video anomaly detection. Converging to such points brings the learning process to an early stop and impedes models from bringing about the expected performance gains of multiple pretext tasks.

In this paper, we propose to sequentially learn multiple pretext tasks for video anomaly detection, and our method is able to provide continual optimization directions, which avoids converging to pareto-optimal points. We arrange the sequential learning order of multiple pretext tasks according to their difficulties in an ascending manner. The difficulty of a task refers to the difficulty of transferring knowledge learned from this task to improve the performance of anomaly detection. Easy tasks usually bring about more performance gains while difficult tasks not. Essentially, our method first relaxes the learning objective of anomaly detection to that of an easy pretext task, and then gradually refines it with more challenging pretext tasks along the learning process. In this way, our model is able to gradually transfer knowledge learned from pretext tasks to model anomalies from coarse to fine, which encourages our model to explore better solutions for video anomaly detection. As shown in Fig. 1, our learning method displays superiority over other sequential learning orders, because they do not consider the difficulties of pretext tasks.

We select three widely-used tasks to model temporal and spatial normality in our learning method. Our method starts with the frame reconstruction task (Rec), then goes on to learn the frame prediction task (Pre) and at last learns the frame order classification task (Cls). We introduce a contrastive loss to push in-order (*i.e.*, positive samples) and out-of-order (*i.e.*, negative samples) inputs apart, which constrains the latent encoding space to achieve better discrimination for models to classify them. We evaluate our methods on three datasets, Avenue [24], ShanghaiTech [27] and UCSD Ped2 [28]. Extensive experiments demonstrate the effectiveness of our method.

In summary, our contributions are two-fold.

- As far as we know, our method is the first attempt to sequentially learn multiple pretext tasks according to their difficulties in an ascending manner, which brings about the expected performance gains in video anomaly detection.
- We introduce a new contrastive loss to constrain the latent space, which grants the trained model with better discrimination for classifying anomalies.

2. Related Work

2.1. Video Anomaly Detection

We review related deep unsupervised anomaly detection methods and divide them into two categories, single pretext task and multiple pretext tasks. Specifically, the term "unsupervised" refers to methods that use only normal data during training. Most unsupervised methods train models to perform a single pretext task with normal samples and assume that the trained models could not perform the pretext task well with abnormal samples. Some works [14, 26, 11, 35, 30, 37, 16] trained models to reconstruct input frames as the pretext task while other works [21, 1] trained models to predict future frames. Although these methods achieved good results, the nonalignment between training tasks (*e.g.*, frame reconstruction, frame prediction) and the testing task (*i.e.*, anomaly detection) still causes a high false alarm rate. Therefore, recent methods[22, 38, 9, 17, 8] began to train models with multiple pretext tasks instead of one single pretext task. Liu *et al.* [22] combined frame prediction and optical-flow image reconstruction as learning objectives. They concatenated two auto-encoders to perform these two tasks. Some other works [9, 17, 8] designed one encoder model with multiple task heads to learn multiple pretext tasks at the same time. Georgescu *et al.* [9] included four tasks (arrow of time, motion irregularity, middle box prediction and model distillation) as pretext tasks. Huang *et al.* [17] introduced contrastive learning methods into anomaly detection and augmented the raw input by shuffling, reversing, rotating and accelerating frames. However, learning multiple pretext tasks to improve the performance of anomaly detection is not fully explored in these works. The reason is that, these methods learn multiple pretext tasks simultaneously, which could easily converge to sub-optimal solutions. Differently, our method designs a tailored sequential learning order, which gradually involves more pretext tasks and sequentially learns them to avoid sub-optimal solutions and achieves better performances.

2.2. Multi-Task Learning

Multi-task learning [39, 45, 40, 42, 5] is adding its appeal to more and more researchers in deep learning. Some methods simultaneously learn multiple tasks by designing complicated network architectures [15] and optimization strategies [7, 34, 41, 12]. These methods are conducted under supervisions and aim to achieve overall performance improvements for multiple tasks. Our goal differs significantly from these methods. We aim to transfer knowledge learned from multiple pretext tasks for modeling normality, so that the performance of anomaly detection is improved. To this end, we propose to sequentially learn multiple pretext tasks according to their difficulties in an ascending manner. Our learning method prevents converging to sub-optimal solutions and ensure the expected performance gains of learning multiple pretext tasks.

Previous work of Pentina *et al.* [32] also proposed to learn multiple tasks sequentially. They arranged the learning order of multiple tasks according to their relatedness so that learning new tasks did not undermine models' performances on previously learned tasks. In this way, they aimed to improve the performances on all tasks. Differently, multiple pretext tasks in our method are arranged according to their difficulties so that knowledge learned from these tasks could better facilitate our model to detect anomalies. The trained model's ability to perform these pretext tasks in [32], however, does not fall into the scope of our concerns. Our sequential learning method would inspire improving the performance of a specific downstream task when learning multiple pretext tasks.

3. Method

3.1. Overview

Video anomaly detection aims to automatically detect anomalous events, which do not conform to human expectations and deviate significantly from normal behavior patterns in videos. Due to the unbounded and rare nature of video anomalies, most works have been directed at unsupervised methods. Unsupervised video anomaly detection can be stated as follows. Given a training set $\mathbb{V}$ consisting of n frames $\{\mathrm{F}_1, \mathrm{F}_2, \ldots, \mathrm{F}_i, \ldots, \mathrm{F}_n\}$ from videos of normal events, the goal is to learn normal patterns and an anomaly scoring function $f : \mathrm{F}_i \rightarrow \mathbb{R}$ during training. During testing, the learned anomaly scoring function assigns large scores to anomaly frames for discrimination.

As is shown in Fig. 2, our method consists of one shared encoder for encoding input snippets as latent codes, a projection block for projecting latent codes and three task heads for learning multiple pretext tasks. We adopt an object detector to construct object-level clips $\mathrm{X} \in \mathbb{R}^{W \times H \times C}$ from each frame F_i and concatenate T consecutive clips in frames $[\mathrm{F}_i, \mathrm{F}_{i+1}, \ldots, \mathrm{F}_{i+T}]$ to obtain the snippet $\mathrm{S} \in \mathbb{R}^{T \times W \times H \times C}$ as the input of our method. During testing, the maximum prediction error of all objects in each frame indicates the frame-level anomaly score, due to the assumption that anomalies are usually unpredictable [21].

3.2. Model Architecture

Our model includes one shared encoder and three task heads for learning multiple pretext tasks. Our encoder $E(\cdot)$ is similar to the one in UNet [23], but does not have any skip connections. The task heads are the reconstruction head $H_{\text{rec}}(\cdot)$, the prediction head $H_{\text{pre}}(\cdot)$ and the classification head $H_{\text{cls}}(\cdot)$. The reconstruction head and the prediction head consist of consecutive deconvolution blocks and they take the latent code, *i.e.*, the output of the shared encoder, as input. The latent code is projected to a new latent space by the projection block $B_{\text{pro}}(\cdot)$ and then fed to the classification head. Both the classification head and the projection block consist of several fully-connected layers.

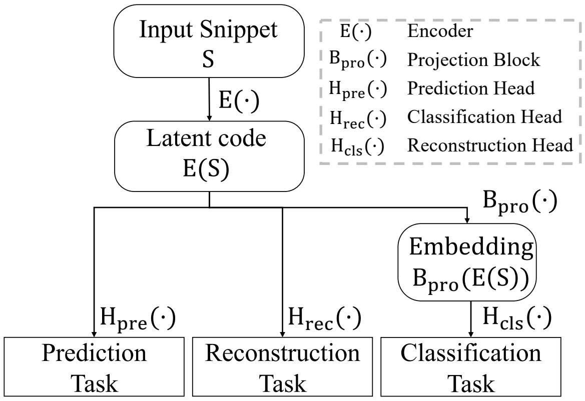


Figure 2. Overview of our method.

3.3. Multiple Pretext Tasks

We select three widely used pretext tasks in our method. These tasks are the frame prediction task, the frame reconstruction task and the frame order classification task. They can be broadly classified into two categories, namely the generative pretext tasks and the discriminative pretext tasks. Generative tasks are the prediction task and reconstruction task. The motivation for choosing these tasks is that most anomalies are visually deviant from normal events, thus less easily to be reconstructed or predicted. The discriminative task is the frame order classification task. The motivation for choosing it is that most anomalies are temporally inconsistent, thus classifying out-of-order frames could help to detect anomalies. We also introduce a contrastive loss to make the representations of learned normality more discriminative. This is helpful to prevent over-generalizing the trained model's ability for some visually indistinct anomalies.

Frame Prediction. Learning the frame prediction task is effective in modeling temporal normality [21, 1]. Most anomalies do not conform to human expectations and are usually deviant from model predictions. Therefore, we believe that the frame prediction task is able to reveal a global picture of optimal solutions for video anomaly detection. We train our model to predict future frames based on a snippet S consisting of four consecutive frames. We draw inspiration from previous tasks of "middle frame prediction", and propose to perform bi-directional frame prediction. We train our model to predict future frames in both directions, forward and backward. We construct a forward snippet $\overrightarrow{\mathrm{S}} = [\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_T]$ and a backward snippet $\overleftarrow{\mathrm{S}} = [\mathrm{X}_{2T+1}, \mathrm{X}_{2T-1}, \ldots, \mathrm{X}_{T+2}]$ as the raw inputs. The object-level X_{T+1} in the middle frame is the prediction target. We use the shared encoder to first encode these two snippets, and use the prediction head to predict X_{T+1}, yielding $H_{\mathrm{pre}}(E(\overrightarrow{\mathrm{S}}))$ and $H_{\mathrm{pre}}(E(\overleftarrow{\mathrm{S}}))$ respectively. The final prediction result $\hat{\mathrm{X}}$ of our model is given by

$$\hat{\mathrm{X}} = \frac{H_{\mathrm{pre}}(E(\overrightarrow{\mathrm{S}})) + H_{\mathrm{pre}}(E(\overleftarrow{\mathrm{S}}))}{2}. \tag{1}$$

We use L2-norm to evaluate the differences between the prediction result and the ground truth, X_{T+1}. The prediction loss $\mathcal{L}_{\mathrm{pre}}$ is given by

$$\mathcal{L}_{\mathrm{pre}} = ||\hat{\mathrm{X}} - \mathrm{X}_{T+1}||_2. \tag{2}$$

In order to generate smooth prediction frames, we also implement a gradient loss. The gradient loss $\mathcal{L}_{\mathrm{grd}}$ is given by

$$\mathcal{L}_{\mathrm{grd}} = ||\nabla\hat{\mathrm{X}} - \nabla\mathrm{X}_{T+1}||_1, \tag{3}$$

where ∇ denotes calculating the gradients. The objective of this task is to minimize both $\mathcal{L}_{\mathrm{pre}}$ and $\mathcal{L}_{\mathrm{grd}}$.

Frame Reconstruction. Learning the frame reconstruction task is effective in modeling spatial normality [14, 26, 11, 30, 16]. This task encourages our model to encode more concrete information about normality, which helps to discriminate against anomalies. We use the shared encoder $E(\cdot)$ to first encode S into the latent space and then reconstruct the input snippet as $H_{\mathrm{rec}}(E(\mathrm{S}))$ from the latent code $E(\mathrm{S})$. We use L2-norm to evaluate the differences between the reconstructed snippet $H_{\mathrm{rec}}(E(\mathrm{S}\ ))$ and the input snippet S. The objective of this task is to minimize the reconstruction loss $\mathcal{L}_{\mathrm{rec}}$, given by

$$\mathcal{L}_{\mathrm{rec}} = ||H_{\mathrm{rec}}(E(\mathrm{S})) - \mathrm{S}||_2. \tag{4}$$

Frame Order Classification. Learning the frame order classification task is effective in modeling temporal consistency of normality [8]. This task is to classify in-order and out-of-order snippets. It encourages our model to learn more discriminative information about normality, which could not be captured by the reconstruction task and the prediction task. As indicated in the work of Chen *et al.* [4], a projection block is able to improve the latent code's quality. We also adopt a projection block, noted as $B_{\mathrm{pro}}(\cdot)$, before the classification head. We generate out-of-order samples by shuffling and accelerating in-order snippets, because they simulate temporal anomalous events in the real world. The shuffled snippet is noted as $\mathrm{S}_{\mathrm{shu}}$. The accelerated snippet is noted as $\mathrm{S}_{\mathrm{acc}}$. We also generate in-order samples by reversing the original snippets, because they preserve the continuity of normal events. The reversed snippet is noted as $\mathrm{S}_{\mathrm{rev}}$. We assign pseudo label 0 to out-of-order snippets and 1 to in-order snippets. The learning objective of the classification head is to correctly classify the original and augmented snippets under the supervision of pseudo labels. The classification loss $\mathcal{L}_{cls}$ takes the form of binary cross entropy, given by

$$\mathcal{L}_{cls} = \begin{array}{l} \log(1-g(\mathrm{S})) + \log(1-g(\mathrm{S}_{\mathrm{acc}})) \\ -\log(g(\mathrm{S}_{\mathrm{rev}})) - \log(g(\mathrm{S}_{\mathrm{shu}})) \end{array}, \tag{5}$$

where,

$$g(\cdot) = H_{\mathrm{cls}}\Big(B_{\mathrm{pro}}\big(E(\cdot)\big)\Big). \tag{6}$$

We further introduce a contrastive loss for making the learned latent codes more discriminative. Contrastive learning leverages the similarities between input samples and augmented ones to learn discriminative representations. Positive samples are encouraged to be crowded together and negative samples pushed apart. By imposing the contrastive learning constraints, the projected latent codes of normal and anomalous events become more distinctive and far apart. Cosine-similarity $\mathrm{sim}(\cdot,\cdot)$ is used to evaluate the similarities between the latent codes of positive samples and negative samples, given by

$$\mathrm{sim}(a,b) = \frac{a^{\mathrm{T}}b}{||a||_2||b||_2}, \tag{7}$$

where a and b denote the latent codes. We use τ to denote the temperature parameter. The raw snippets and reversed snippets are treated as positive samples and the shuffled snippets and the accelerated snippets as negative samples. The similarity score s_{pos} between positive samples is given by

$$s_{\mathrm{pos}} = \sum \exp(\frac{\mathrm{sim}(E(\mathrm{S}), E(\mathrm{S}_{\mathrm{rev}}))}{\tau}). \tag{8}$$

The similarity score s_{neg} between negative and positive samples is given by

$$\begin{aligned} s_{\mathrm{neg}} = &\sum \exp(\frac{\mathrm{sim}(E(\mathrm{S}), E(\mathrm{S}_{\mathrm{shu}}))}{\tau}) \\ &+ \sum \exp(\frac{\mathrm{sim}(E(\mathrm{S}), E(\mathrm{S}_{\mathrm{acc}}))}{\tau}). \end{aligned} \tag{9}$$

The contrastive learning loss $\mathcal{L}_{con}$ is given by

$$\mathcal{L}_{con} = -\log\frac{s_{\mathrm{pos}}}{s_{\mathrm{pos}} + s_{\mathrm{neg}}}. \tag{10}$$

The objective of this task is to minimize the classification loss $\mathcal{L}_{cls}$ and the contrastive loss $\mathcal{L}_{con}$.

3.4. Sequential Learning Order

We observe that the conventional learning method, *i.e.*, simultaneously learning multiple pretext tasks, is prone to sub-optimal solutions for video anomaly detection. Some previous works [8, 17] tried to alleviate this problem by assigning different weights to multiple pretext tasks. However, models still risk converging to pareto-optimal points,

Tasks\Phase	**Phase1**	**Phase2**	**Phase3**
Prediction Task	✓	✓	✓
Reconstruction Task		✓	✓
Classification Task			✓

Table 1. The sequential learning order of our method.

rendering less satisfying performances. The possible reason is that these tasks could have potentially conflicting objectives and simultaneous learning encourages an early stop of model optimization. Inspired by the learning process of human, we propose to sequentially learn multiple pretext tasks according to their difficulties from easy to hard. In this way, the proposed learning method is able to provide our model with constantly evolving learning objectives and prevent our model from converging to sub-optimal solutions.

Task Difficulty Measurement. The difficulty of a task refers to the difficulty of transferring knowledge learned from this task to improve the performance of anomaly detection. And it is measured by the objective AUC performance. The frame prediction task is easier because it targets unpredictable anomalies, which usually deviate from model expectations and are the most frequent and common anomalies. And learning this task alone from scratch renders a better performance in detecting anomalies. The classification task is more difficult because it targets anomalies in the latent space, which are less frequent and indistinctive anomalies. And learning this task alone from scratch renders a worse performance in detecting anomalies. Easy pretext tasks and difficult pretext tasks are complimentary, because each pretext task is designed to model some characteristics of anomalies by transferring knowledge learned from this task. Easy pretext tasks model more general and common characteristics (*i.e.*, anomalies in coarse-scale) and difficult pretext tasks model more specific and rare characteristics (*i.e.*, anomalies in fine-scale). Therefore, we design a new learning method which sequentially learns all these tasks.

Sequential Learning Order Design. The sequential learning order is shown in Tab. 1 and the difficulties of these tasks are used as a criterion for deciding the learning order. The optimal order refers to arranging multiple pretext tasks according to their difficulties in an ascending manner. It consists of three phases. (1) In phase one, our model learns to perform the frame prediction task. This task is relatively easy for models to start with and models could learn a better starting point for other more challenging tasks. (2) In phase two, our model learns to perform the frame prediction task and the frame reconstruction task together. The prediction head and the reconstruction head share the same encoder from phase one. In this way, the reconstruction head is able to leverage knowledge learned from the previous task. (3) In phase three, we train the classification head to classify the time order of input snippets. The classification head is trained to discriminate latent codes of in-order (consecutive) and out-of-order (shuffled/accelerated) input snippets. The criterion for deciding the learning order is objective so that more pretext tasks can be easily inserted into the current learning order according to their difficulties.

3.5. Anomaly Detection

During testing, we first calculate the prediction error $\mathcal{L}(\mathrm{S})$ of every snippets S per frame, given by

$$\mathcal{L}(\mathrm{S}) = ||H_{\mathrm{pre}}(E(\mathrm{S})) - \mathrm{X}_{T+1}||_2. \tag{11}$$

We then assign 0 as anomaly score to any frame without detected salient objects. For a frame $\mathrm{F_i}$ with m salient objects, the maximum anomaly score $\mathcal{L}_{\max}$ among these objects is given by

$$\mathcal{L}_{\max} = \max\{\mathcal{L}(\mathrm{S}_1), \mathcal{L}(\mathrm{S}_2), \ldots, \mathcal{L}(\mathrm{S}_m)\}. \tag{12}$$

The learned anomaly scoring function is given by

$$f = \begin{cases} 0 & \mathrm{F_i} \text{ has } 0 \text{ objects} \\ \mathcal{L}_{\max} & \mathrm{F_i} \text{ has } m \text{ objects} \end{cases}. \tag{13}$$

The frame-level score is further smoothed by a median filter whose window size is 17 to ensure the temporal consistency of videos.

4. Experiments

4.1. Experimental Setting

Datasets. We evaluate our method on three challenging datasets, namely Avenue [24], ShanghaiTech [27] and UCSD Ped2 [28]. Each dataset can be divided into training sets and testing sets. Only normal frames are included in the training sets. (1) Avenue [24]: A total of 16 training videos and 21 testing videos are included in the Avenue dataset. The anomalous events in the testing sets include anomalous pedestrian movements, wrong directions, and anomalous objects. (2) ShanghaiTech [27]: A total of 13 different surveillance scenes are included in the ShanghaiTech dataset. It is one of the most challenging video anomaly detection datasets. The testing sets contain over 130 anomalous events. (3) UCSD Ped2 [28]: A total of 36 training videos and 12 testing videos are included in the Ped2 dataset. The Ped2 dataset uses video footage captured by cameras fixed at high elevations overlooking the sidewalk. Anomalies in the dataset are mainly caused by the presence of non-human entities (cars, wheelchairs, skateboards, bicycles, etc.) as well as some abnormal behavior patterns of pedestrians.

$\mathcal{L}_{\text{pre}}$	$\mathcal{L}_{\text{grd}}$	$\mathcal{L}_{\text{rec}}$	$\mathcal{L}_{\text{sps}}$	$\mathcal{L}_{\text{cls}}$	$\mathcal{L}_{\text{con}}$	Avenue	Ped2
✓	✓	✓	✓	✓	✓	76.51	83.85
✓	✓	✓	✓	✓	-	81.84	76.50
-	-	✓	✓	✓	✓	**84.47**	63.76
✓	✓	-	-	✓	✓	78.58	88.40
-	-	✓	✓	-	-	73.32	82.69
✓	✓	-	-	-	-	82.10	72.04
✓	-	-	-	-	-	79.15	**92.29**

Table 2. AUC (%) performances of models trained with different loss combinations in Avenue and Ped2 datasets. We take six losses into consideration, namely prediction loss $\mathcal{L}_{\text{pre}}$, gradient loss $\mathcal{L}_{\text{grd}}$, reconstruction loss $\mathcal{L}_{\text{rec}}$, sparsity loss $\mathcal{L}_{\text{sps}}$, classification loss $\mathcal{L}_{\text{cls}}$ and contrastive loss $\mathcal{L}_{\text{con}}$.

Evaluation Metric. We evaluate our method on a frame-level metric. Following previous popular works [14, 11, 25, 36], we adopt the area under the receiver operating characteristic curve (AUC) as the evaluation metric. A higher AUC indicates a better video anomaly detection performance.

Compared Methods. For single pretext task based methods, we compare our method with reconstruction based methods like Conv-AE [14], ConvLSTM-AE [26], MNAD-R [30] and prediction based methods like Frame-Pred [25], MNAD-P [30] and VEC [44]. For multiple pretext tasks based methods, we compare our method with MemAE [11], ST-AE [46], AMC [29], GMFC-VAE [6], AnoPCN [43], HF2VAD [22], AD-Con [17] and object-centric [8].

4.2. Implementation Details

The encoder consists of a convolution layer and three convolution blocks. The prediction head and the reconstruction head are symmetrical to the encoder, consisting of three de-convolution blocks and a convolution layer at the end. The prediction head and the reconstruction head have the same network architecture for the first three blocks. The difference lies in the last convolution layer. For smaller datasets like Avenue and Ped2, we make a minor revision to the auto-encoder and remove the last block to make the network smaller. The detailed architectures can be found in ***supplementary materials***. We train our model on salient objects extracted from video frames. Following the work of Liu *et al.* [22], we first preprocess the three datasets by applying an object detector. We use pretrained fast RCNN [10] as the object detector in our method. The objects of interest are uniformly scaled to 32 × 32. We stack four consecutive clips to form a snippet, which is the raw input snippet S of our model. During testing, we treat the maximum score of all objects in one frame as the frame-level anomaly score. We use PyTorch [31] to implement our method and adopt Adam optimizer [20] to optimize it. The training process follows the sequential learning order described in Tab. 1. The batchsize is fixed at 256. We conduct experiments on an NVIDIA GeForce GTX 1080 Ti and an Intel(R) Core(TM) i7-7800X CPU @ 3.50GHz. The training time of each phase for ShanghaiTech, Avenue and Ped2 is roughly 48 hours, 12 hours and 2 hours.

4.3. Experiments on Multi-Task Learning

We conduct the following experiments to show that simultaneously optimizing multiple objectives could converge to sub-optimal solutions. We adopt an auto-encoder architecture but slightly modify it by adding extra memory modules in the work of Gong *et al.* [11]. We use the memory modules to constrain the auto-encoder's generalization ability to prevent the auto-encoder from wrongly reconstructing anomalies. To this end, we use six loss functions for optimization, including the $\mathcal{L}_{\text{pre}}, \mathcal{L}_{\text{grd}}, \mathcal{L}_{\text{rec}}, \mathcal{L}_{\text{cls}}, \mathcal{L}_{\text{con}}$ in Eqs. (2) to (5) and (10), as well as a memory sparsity loss $\mathcal{L}_{\text{sps}}$. The sparsity loss constrains the memory addressing variable to be sparse enough, thus using as few memory items as possible for reconstruction. Please refer to [11] for the details of the sparsity loss function $\mathcal{L}_{\text{sps}}$. We use memory modules in this experiment, but not in the final experiment. The reason is that the memory modules are effective in more complex auto-encoders, *i.e.*, auto-encoders with skip connections, while the architecture of our auto-encoder is fairly simple. This means that adding memory modules poses such a strict constraints that our auto-encoder could no longer faithfully reconstruct normal snippets.

We train seven models with sum of different loss functions for 80 epochs in two datasets, Avenue and Ped2, and record the best performance for comparison. The complete results are reported in Tab. 2. Results further demonstrate that training models simultaneously with multiple objectives fails to bring about the expected performance gains and could cause severe performance drops. For example, we achieve the best AUC performance in Ped2 with only $\mathcal{L}_{\text{pre}}$, surpassing the model trained with all losses. The reason is that models could get stuck in a trade-off solution among all objectives, where optimizing any of them undermines the rest of the objectives. Such solutions are not always good solutions for video anomaly detection.

4.4. Comparisons with State-of-the-art Methods

Qualitative Results. In Fig. 3, we visualize some frame prediction results compared to the state-of-the-art method [22]. It shows that our method causes larger prediction errors even in the smooth background, which means that our method is more discriminative against video anomalies.

In Fig. 4, we visualize the predicted anomaly scores from each phase and the anomaly ground truth duration in several videos. It shows that the predicted anomaly scores match better with the ground truth after every phase.

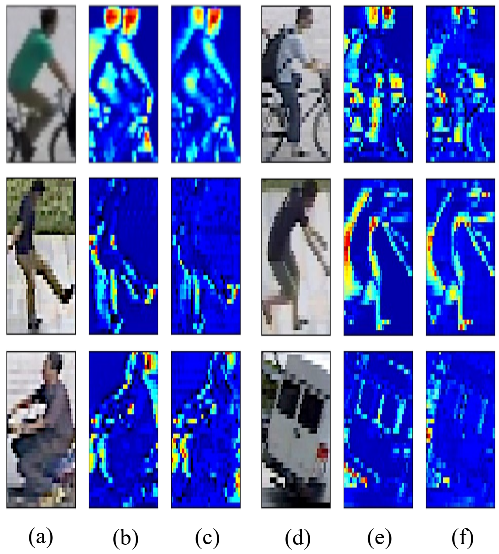


Figure 3. Comparisons of frame prediction results between HF2VAD [22] and our method. From left to right, we show the ground truth clips in column (a) and (d), difference maps of our method in column (b) and (e), difference maps of HF2VAD in column (c) and (f). From top to bottom, we show three different kinds of anomalies, namely riding bicycles, irregular motions and vehicles.

Model	Task	AUC		
		Ped2	Avenue	ShTech
Conv-AE [14]	single	90.0	70.2	-
ConvLSTM-AE [26]	single	88.1	77.0	-
CONV-VRNN [25]	single	96.1	85.8	-
ST-AE [46]	single	91.2	80.9	-
MNAD-R [30]	single	90.2	82.8	69.8
MNAD-P [30]	single	97.0	88.5	70.5
MemAE [11]	multiple	94.1	83.3	71.2
AD-Con [17]	multiple	**98.1**	88.8	77.2
Ours	multiple	97.6	**90.9**	**78.8**

Table 3. Comparisons of AUC (%) performances with state-of-the-art methods that only use RGB images as inputs.

Quantitative Results. We compare our method with state-of-the-art methods that only use RGB images as inputs in Tab. 3. We do not compare our method with top-performing methods that use extra data [33] or extra input features (e.g., optical-flow images) from pre-trained models [2] for fair comparison. The performances of the compared methods are taken from their original papers. From Tab. 3, it can be seen that: (1) Our method outperforms all single-pretext task based methods. This demonstrates that sequentially learning more pretext tasks can alleviate the nonalignment problem. (2) Our method achieves state-of-the-art performances on both the Avenue and ShanghaiTech datasets, gaining improvements of 2.1% and 1.6% respectively compared with the state-of-the-art method [17]. The superior results demonstrate the effectiveness of sequentially learning multiple pretext tasks. (3) On the Ped2 dataset, our method performs slightly worse than the work of Huang *et al.* [17]. The probable reason is that the object detector [10] in our method fails to provide good object detection results in crowded scenes from the Ped2 dataset. The method in [17] uses frame-level inputs, which favor anomaly detection in crowded scenes. We show some failure object detection results on the ***supplementary materials***. We will address anomaly detection in crowded scenes by introducing new scene-level pretext tasks in our future work.

Model	Optical-flow	AUC		
		Ped2	Avenue	ShTech
AMC [29]	✓	96.2	86.9	-
GMFC-VAE [6]	✓	92.2	83.4	-
VEC [44]	✓	97.3	90.2	74.8
AnoPCN [43]	✓	96.8	86.2	73.6
Frame-Pred [21]	✓	95.4	85.1	72.8
object-centric [19]	✓	94.3	87.4	78.7
STCEN [13]	✓	96.9	86.6	73.8
BDPN [3]	✓	98.3	90.3	78.1
AMSRC [18]	✓	99.3	**93.8**	76.3
MSTL [8]	✓	97.6	91.5	**82.4**
HF2VAD [22]	✓	**99.3**	91.1	76.2
Ours	✗	97.6	90.9	78.8

Table 4. Comparisons of AUC (%) performances with state-of-the-art methods that use both RGB images and optical-flow images as inputs.

We further compare our method with state-of-the-art methods that use both RGB images and optical-flow images as inputs in Tab. 4. From Tab. 4, it can be seen that: (1) Without using optical-flow images as inputs, our method still achieves competitive performances on the ShanghaiTech dataset. The work of [8] performed better than ours. This could be attributed to the fact they use optical-flow images as inputs and more pretext tasks (four in theirs versus three in ours) (2) Our method performs worse than the work [22] on the Ped2 dataset and the work [22] on the Avenue dataset, due to the lack of using optical-flow images to help discriminate motion-relevant anomalies. The improvements of our method with optical-flow images as inputs are shown in ***supplementary materials***.

4.5. Ablation Study

Ablation Study on Learning Orders. We conduct the following experiments to see the extent to which the performance of a model is influenced by the learning order of pretext tasks. We retrain our model to learn three pretext tasks in different orders and the results are reported in Tab. 5. The abbreviations for the prediction task, the reconstruction task

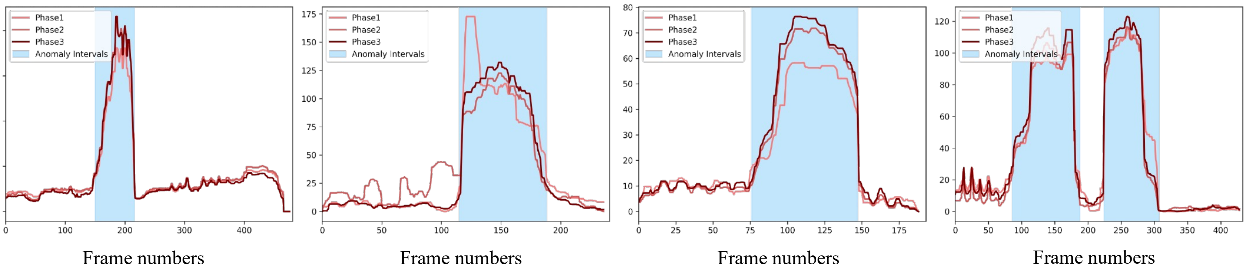


Figure 4. Curves of predicted anomaly scores from each phase. Frames in white and colored windows are the ground truth normal and anomaly frames, respectively. The anomaly score curves match better with the ground truth after sequentially learning multiple pretext tasks.

Dataset	Sequential						Simultaneous
	Learning Order			AUC			AUC
	Phase1	Phase2	Phase3	Phase1	Phase2	Phase3	
Ped2	Pre	+Rec	+Cls	91.34	95.17	**97.57**	87.97
		+Cls	+Rec		93.44	95.66	91.46
	Rec	+Pre	+Cls	73.12	93.01	94.18	88.04
		+Cls	+Pre		88.34	92.02	90.85
	Cls	+Pre	+Rec	-	94.29	95.31	87.53
		+Rec	+Pre		59.73	93.41	91.41
Avenue	Pre	+Rec	+Cls	89.61	90.57	**90.85**	80.21
		+Cls	+Rec		83.24	82.40	76.11
	Rec	+Pre	+Cls	83.34	87.47	86.71	75.70
		+Cls	+Pre		79.82	78.22	77.86
	Cls	+Pre	+Rec	-	76.53	75.53	74.89
		+Rec	+Pre		77.97	75.51	78.29

Table 5. AUC (%) performances of models trained sequentially with different learning orders, and models trained simultaneously with different weight assignments in Avenue and Ped2 datasets. The results show that our sequential learning method is able to achieve the most performance gains when learning multiple pretext tasks.

and the classification task are Pre, Rec and Cls respectively. The sequential learning order is divided into three different phases, and in each new phase, a new task is added to the learning objective. Switching the learning orders of pretext tasks renders sub-optimal solutions, *e.g.*, from 97.57% ("Pre-Rec-Cls") to 92.02% ("Rec-Cls-Pre") in Ped2 and from 90.85% ("Pre-Rec-Cls") to 75.51% ("Cls-Rec-Pre") in Avenue. Our proposed learning order, *i.e.*, "Pre-Rec-Cls", brought consistent performance improvements in both small (Ped2) and large (Avenue) datasets, while other learning orders not. It shows that learning multiple pretext tasks sequentially, especially from easy to difficult, is better for video anomaly detection.

Ablation Study on Re-weighting. We conduct the following experiments to validate that our sequential learning method can not be replaced by re-weighting multiple pretext tasks. We convert each sequential learning order into its simultaneous learning "equivalence" by re-weighting them based on the length of training time. For example, the learning order of "Pre-Rec-Cls" is converted to weights of "3:2:1" for three pretext tasks respectively. The results are reported in the last column of Tab. 5. Switching the sequential learning method to simultaneous learning method incurred sharp performance drops, *e.g.*, from 97.57% ("Cls-Rec-Pre") to 87.97% in Ped2 and from 86.71% ("Rec-Pre-Cls") to 75.70% in Avenue. Most sequential learning method outperformed their simultaneously learning counterpart, except from the reversed learning order in the last row. The possible reason is that learning multiple pretext tasks from difficult to easy is not ideal. Overall, the results of this ablation study shows that simply assigning different weights to different pretext tasks is suboptimal compared to learning them sequentially.

Results of the aforementioned two experiments in ShanghaiTech dataset can be found in the ***supplementary materials*** .

5. Conclusion and Future Work

In this paper, we propose to sequentially learn multiple pretext tasks for video anomaly detection. The sequential learning order of multiple pretext tasks follows their difficulties in an ascending manner. Our method gradually involves more challenging pretext tasks so that the learning objective is constantly evolving. By doing so, our method can avoid converging to sub-optimal solutions. Besides, we introduce a new contrastive loss to the classification task. The contrastive loss can make the learned representations of normality more discriminative by posing constraints on the latent space. The contrastive loss pushes normal samples and pseudo-abnormal samples apart. Experiment results on three datasets can demonstrate the effectiveness of our method.

Our method ignores the scene information and solely relies on object detectors to detect salient objects, which fails to detect anomalies that are context-related in crowded scenes (*e.g.*, the scenes in Ped2). The scene information plays an important role in defining video anomalies. For example, car driving in the pedestrian is deemed as anomalies, while a car on the road not. In future work, we will design a frame-level pretext task to learn the scene information. The pretext task measurement could be susceptible to the influences of different models and datasets. In future work, we will design a more systematic and better ways to define the difficulty of each pretext tasks.

Acknowledgments This work was supported by the Natural Science Foundation of China (NSFC) under Grants No. 62172041 and No. 62176021.

References

[1] Marcella Astrid, Muhammad Zaigham Zaheer, and Seung-Ik Lee. Synthetic temporal anomaly guided end-to-end video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 207–214, 2021. 2, 3

[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*, page 103656, 2023. 7

[3] Chengwei Chen, Yuan Xie, Shaohui Lin, Angela Yao, Guannan Jiang, Wei Zhang, Yanyun Qu, Ruizhi Qiao, Bo Ren, and Lizhuang Ma. Comprehensive regularization in a bi-directional predictive network for video anomaly detection. In *Proceedings of the American Association for Artificial Intelligence*, pages 1–9, 2022. 7

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 4

[5] Michael Crawshaw. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*, 2020. 2

[6] Yaxiang Fan, Gongjian Wen, Deren Li, Shaohua Qiu, Martin D Levine, and Fei Xiao. Video anomaly detection and localization via gaussian mixture fully convolutional variational autoencoder. *Computer Vision and Image Understanding*, 195:102920, 2020. 6, 7

[7] Heshan Fernando, Han Shen, Miao Liu, Subhajit Chaudhury, Keerthiram Murugesan, and Tianyi Chen. Mitigating gradient bias in multi-objective learning: A provably convergent stochastic approach. *arXiv preprint arXiv:2210.12624*, 2022. 1, 2

[8] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12742–12752, 2021. 2, 4, 6, 7

[9] Mariana-Iuliana Georgescu, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Claudiu Popescu, and Mubarak Shah. A background-agnostic framework with adversarial training for abnormal event detection in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:4505–4523, 2022. 1, 2

[10] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015. 6, 7

[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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1705–1714, 2019. 2, 4, 6, 7

[12] Michelle Guo, Albert Haque, De-An Huang, Serena Yeung, and Li Fei-Fei. Dynamic task prioritization for multitask learning. In *Proceedings of the European Conference on Computer Vision*, pages 270–287, 2018. 1, 2

[13] Yi Hao, Jie Li, Nannan Wang, Xiaoyu Wang, and Xinbo Gao. Spatiotemporal consistency-enhanced network for video anomaly detection. *Pattern Recognition*, 121:108232, 2022. 7

[14] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 733–742, 2016. 1, 2, 4, 6, 7

[15] Nicolas Heess, Greg Wayne, Yuval Tassa, Timothy Lillicrap, Martin Riedmiller, and David Silver. Learning and transfer of modulated locomotor controllers. *arXiv preprint arXiv:1610.05182*, 2016. 2

[16] Jinlei Hou, Yingying Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, and Hong Zhou. Divide-and-assemble: Learning block-wise memory for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8791–8800, 2021. 2, 4

[17] Chao Huang, Zhihao Wu, Jie Wen, Yong Xu, Qiuping Jiang, and Yaowei Wang. Abnormal event detection using deep contrastive learning for intelligent video surveillance system. *IEEE Transactions on Industrial Informatics*, 2021. 1, 2, 4, 6, 7

[18] Xiangyu Huang, Caidan Zhao, Yilin Wang, and Zhiqiang Wu. A video anomaly detection framework based on appearance-motion semantics representation consistency. *arXiv preprint arXiv:2204.04151*, 2022. 7

[19] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7842–7851, 2019. 7

[20] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 6

[21] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6536–6545, 2018. 1, 2, 3, 7

[22] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13588–13597, 2021. 2, 6, 7

[23] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 3

[24] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2720–2727, 2013. 1, 2, 5

[25] Yiwei Lu, K Mahesh Kumar, Seyed shahabeddin Nabavi, and Yang Wang. Future frame prediction using convolutional vrnn for anomaly detection. In *IEEE International Conference on advanced Video and Signal based Surveillance*, pages 1–8. IEEE, 2019. 6, 7

[26] Weixin Luo, Wen Liu, and Shenghua Gao. Remembering history with convolutional lstm for anomaly detection. In *IEEE International Conference on Multimedia and Expo*, pages 439–444, 2017. 2, 4, 6, 7

[27] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 341–349, 2017. 2, 5

[28] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *IEEE computer society Conference on Computer Vision and Pattern Recognition*, pages 1975–1981, 2010. 1, 2, 5

[29] Trong-Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *Proceedings of the IEEE/CVF international conference on Computer Vision*, pages 1273–1283, 2019. 6, 7

[30] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14372–14381, 2020. 2, 4, 6, 7

[31] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[32] Anastasia Pentina, Viktoriia Sharmanska, and Christoph H Lampert. Curriculum learning of multiple tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5492–5500, 2015. 2, 3

[33] Tal Reiss and Yedid Hoshen. Attribute-based representations for accurate and interpretable video anomaly detection. *arXiv preprint arXiv:2212.00789*, 2022. 7

[34] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in Neural Information Processing Systems*, 31, 2018. 1, 2

[35] Hao Song, Che Sun, Xinxiao Wu, Mei Chen, and Yunde Jia. Learning normal patterns via adversarial attention-based autoencoder for abnormal event detection in videos. *IEEE Transactions on Multimedia*, 22(8):2138–2148, 2019. 2

[36] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, 2018. 6

[37] Che Sun, Yunde Jia, Hao Song, and Yuwei Wu. Adversarial 3d convolutional auto-encoder for abnormal event detection in videos. *IEEE Transactions on Multimedia*, 23:3292–3305, 2020. 2

[38] Che Sun, Chenrui Shi, Yunde Jia, and Yuwei Wu. Learning event-relevant factors for video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2384–2392, 2023. 2

[39] Kai Sun, Richong Zhang, Samuel Mensah, Yongyi Mao, and Xudong Liu. Progressive multi-task learning with controlled information flow for joint entity and relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13851–13859, 2021. 2

[40] Kim-Han Thung and Chong-Yaw Wee. A brief review on multi-task learning. *Multimedia Tools and Applications*, 77(22):29705–29725, 2018. 2

[41] Ben Tu, Axel Gandy, Nikolas Kantas, and Behrang Shafei. Joint entropy search for multi-objective bayesian optimization. *arXiv preprint arXiv:2210.02905*, 2022. 2

[42] Partoo Vafaeikia, Khashayar Namdar, and Farzad Khalvati. A brief review of deep multi-task learning and auxiliary task learning. *arXiv preprint arXiv:2007.01126*, 2020. 2

[43] Muchao Ye, Xiaojiang Peng, Weihao Gan, Wei Wu, and Yu Qiao. Anopcn: Video anomaly detection via deep predictive coding network. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1805–1813, 2019. 6, 7

[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 *Proceedings of the 28th ACM International Conference on Multimedia*, pages 583–591, 2020. 6, 7

[45] Yu Zhang and Qiang Yang. An overview of multi-task learning. *National Science Review*, 5(1):30–43, 2018. 2

[46] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua. Spatio-temporal autoencoder for video anomaly detection. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1933–1941, 2017. 6, 7