Segment-aware Evidential Fusion Network for Trustworthy Video Sewer Defect Classification

Chenyang Zhao[1], Chuanfei Hu[1], Jin Deng[1]
[1]Southeast University, Nanjing, China

***Abstract*—An automatic vision-based sewer inspection plays a vital role of sewage system in a modern city. Although the acceptable performances of sewer defect classification are achieved, there is still a gap between the emerged methods and actual application scenarios. The first issue is that the multi-segment information is ignored to represent the sewer defect, resulting in capturing the multi-scale complementarity among segments inefficiently. Second, the inherent uncertainty of sewer defect is not considered, while the serious unknown sewer defect categories would be missed, resulting in the untrustworthy sewer inspection. In this paper, we focus on quick-view (QV)-based sewer inspection, while a segment-aware evidential fusion network (SEFN) is proposed, jointly combining multi-label classification and uncertainty estimation. Specifically, multi-segment splitting module (MSM) is designed based on dynamic information to split the QV sewer video into global-scale and local-scale segments, where the multi-segment can be modeled to represent the multi-scale information of sewer defect. Then, evidential deep learning (EDL) is introduced to quantify the uncertainty, while Dempster-Shafer theory (DST)-based combination rule is utilized to aggregate the expert opinions of multi-scale segments. Furthermore, evidence selecting scheme (ESS) is proposed to alleviate the ambiguity of uncertainty estimation. Extensive experiments are conducted on VideoPipe, in which the superiority of SEFN is demonstrated compared with the state-of-the-art methods.**

I. Introduction

Underground sewerage system is an essential part of modern urban public infrastructure [1], which prevents the secondary pollution of water resources caused by untreated sewage, and contributes to nutrient recovery and replenishment of groundwater supply. However, defects in sewage pipes due to unforeseen factors are inevitable, such as corrosion and wear. These defects would directly cause the functional failures of the underground sewerage system and even result in serious incidents, such as environmental damage [2] and road collapses [3]. Therefore, an automatic inspection of sewer pipes and accurate assessment of defects are vital to the underground sewerage system, which has attracted constant attention and recognition from both academic and industrial communities.

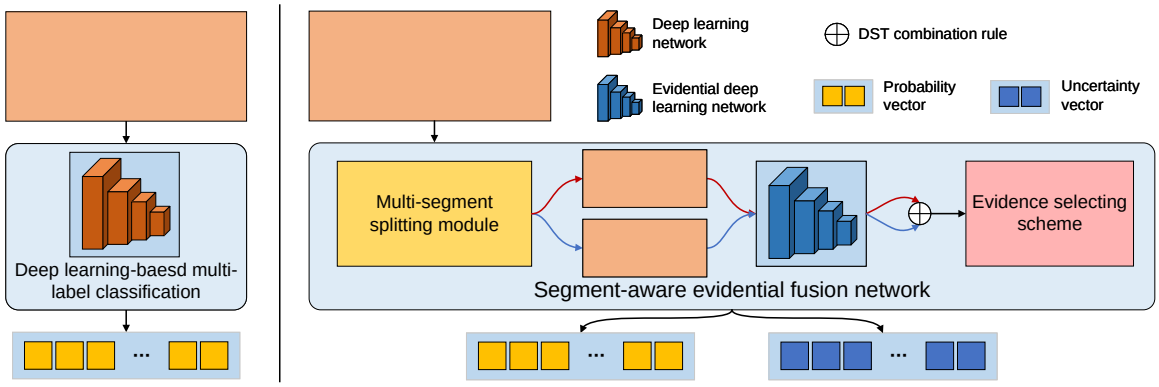
In the past few decades, the vision-based inspection methods have emerged as the mainstream technology [4], where quick-view (QV) [5] and closed-circuit television (CCTV) [6] are the dominant data acquisition modes. Due to the portability and simple operation of QV device, the holistic condition of sewer pipes can be collected in a shorter period of time. Meanwhile, the remote-controlled robot vehicle would be difficult to pass through the sewer pipes with high water level and obstacles [7]. Thus, the QV-based inspection methods yield higher working efficiency compared with the CCTV-based. Here, we focus on the QV-based acquisition mode, exploring an automatic sewer inspection method.

Recent years have witnessed the success of deep learning in industrial applications [9], [12]–[14]. For vision-based sewer inspection system, the emerging deep learning-based methods [16]–[18] can be categorized into sewer defect classification, detection, and segmentation. Since the QV-based sewer inspection is not sensitive to the localization of defects, we focus on the sewer defect classification with multi-label setting. The multi-label setting means that more than one sewer defect category would exist in a QV sewer video, which is closer to the practical scenario of sewerage system assessment [17], [19]. Although these deep learning-based models have achieved satisfactory performances of sewer defect classification, there are two issues should be considered.

*"**First**, the multi-segment information of sewer video is ignored to represent the sewer defects, resulting in capturing the complementarity among the segments inefficiently."* Intuitively, we would observe an object via a global-to-local view. Such procedure of multi-scale view is also utilized by inspectors, since the complementarity information is beneficial for assessing the appearances of sewer pipes.

*"**Second**, the existing deep learning-based models might not focus on the inherent uncertainty of sewer defect in real-world applications [23] comprehensively."* For instance, in the out-of-distribution (OOD) scenario, some categories of sewer defects have not recorded in the historical database, while the trained model has not observed these unknown defect categories. Such untrustworthy results would result in the missed detection and false negative case. Most recently, evidential deep learning (EDL) [24] has become a popular method to uncertainty estimation. Although the availability of EDL for multi-label sewer defect classification has been validated [19], the ambiguity among different sewer categories would still degenerate the performance of uncertainty estimation.

To address these issues, we propose a segment-aware evidential fusion network (SEFN) for QV-based sewer inspection, jointly combining multi-label classification and uncertainty estimation as a unified framework. Specifically, we first design a multi-segment splitting module (MSM) to split the sewer video into global-scale and local-scale segments, where the dynamic information is estimated to capture the moment of scale variation. Then, EDL is introduced to conduct a deep learning-based network for the multi-scale segments, casting the multi-label classification as an uncertainty estimation. Since EDL-based uncertainty of multi-scale segments is quantified independently, we conduct the Dempster-Shafer theory [26] (DST)-based combination rule to aggregate the expert opinions of multi-scale segments. Moreover, evidence selecting scheme (ESS) is proposed to mitigate the ambiguity of uncertainty



(a) Existing paradigm. (b) The proposed segment-aware evidential fusion network (SEFN).

Fig. 1. (a) The multi-segment information of sewer video is ignored in the existing paradigm. (b) Overall of segment-aware evidential fusion network (SEFN) for multi-label sewer defect classification on QV sewer video. The QV sewer video is split into the global-scale and local-scale segments via multi-segment splitting module (MSM), where the optical flow is introduced as the dynamic information to capture the moment of scale-variant. Then, a shared weight evidential deep learning network is constructed to generate the expert opinions independently, while the the expert opinions of multi-scale segments are fused via DST-based combination rule. Furthermore, evidence selecting scheme (ESS) would mitigate the ambiguity of uncertainty estimation. The final probability vector and uncertainty vector are the trustworthy results for the multi-label sewer defect classification.

estimation. The main contributions are summarized as follows:

1) Segment-aware evidential fusion network (SEFN) is proposed for sewer defect classification, jointly combining multi-label classification and uncertainty estimation as a unified framework. To our best knowledge, it is among the first sewer defect classification method to model the multi-scale segments from QV sewer video.
2) Multi-segment splitting module (MSM) based on dynamic information is designed to split the QV sewer video into global-scale and local-scale segments, capturing the natural multi-scale characteristic of sewer video.
3) Evidence selecting scheme (ESS) is proposed to mitigate the ambiguity, strengthening the reliability of the evidence-based uncertainty estimation.

II. Related Work

A. Automated Sewer Inspection

With the rapid development of computer vision in the industrial field, deep learning-based sewer inspection methods have emerged [5], [16], [18], [30]. However, these methods mentioned above belong to the image-based method with intricate preprocessing and postprocessing, while the temporal contextual semantic information among video frames is ignored. There have been the limited exploration on video-based sewer inspection in previous works. Recently, Liu *et al*. [27] release a large scale real-world framework sewer inspection, including two new industrial video datasets, namely QV-Pipe and CCTV-Pipe, which are utilized for multi-label video defect classification and temporal defect localization, respectively. Although the acceptable performances are achieved in the reported methods, the representation of multi-focus characteristic and inherent uncertainty is still an in-depth challenge.

B. Multi-label Video Classification

Since there have not multi-label classification method based on videos for sewer inspection, we explore the relevant methods in the field of multi-label video action classification [31]–[34]. These methods designed specifically for multi-label action recognition might not be effective, since there is a domain gap of contextual semantic information between the human actions and sewer defects. Thus, we argue that it is valuable to explore the domain-specific method for the multi-label sewer defect classification based on videos.

C. Out-of-Distribution Detection

The goal of OOD detection is to distinguish the samples from the unknown category, which have never appeared in the training set. The defect detection methods based on softmax layer often lead to overconfident classification results [35], which are unreliable towards the unknown defect categories. Therefore, it is necessary to estimate the uncertainty of prediction results. Most recently, there have been numerous efforts to avoid such repeated inferences of conventional uncertainty evaluation methods [36]–[38], and instead quantify the uncertainty estimation through model-based approaches [19], [24], [39], [40]. Different from these previous works, we focus on the EDL-based method, and reveal the ambiguity of uncertainty estimation in the multi-label sewer defect classification, which would dramatically weaken the reliability of uncertainty estimation.

III. Methodology

A. Overview

The overall framework of SEFN is shown in Fig. 1, which consists of multi-segment splitting module (MSM), evidential deep learning network, and evidence selecting scheme (ESS).

Given a QV sewer video $\mathbf{X} \in \mathbf{R}^{C\times H\times W\times T}$ with multi-label annotation $\mathbf{y} \in \mathbf{H}^K$. C, H, W, and T denote the number of color channels, height, width, and frames, K denotes the number of defect categories, $\mathbf{H}$ is Hamming space. Specifically, MSM is first conducted to split the $\mathbf{X}$ into the dual-segment as follows:

$$X^G, X^L = \text{MSM}\left(X\right), \tag{1}$$

where X^G and X^L are global-scale and local-scale segments, respectively. Then, an evidential deep learning network E is constructed to quantify the uncertainty as follows:

$$\mathcal{M}^G = E(X^G), \quad \mathcal{M}^L = E(X^L), \tag{2}$$

where the expert opinions $(\mathcal{M}^G, \mathcal{M}^L)$ of multi-scale segments can be aggregated via DST combination rule $\oplus$ as follows:

$$\mathcal{M}^F = \mathcal{M}^G \oplus \mathcal{M}^L, \tag{3}$$

where the derivation of probability $\mathbf{p}$ of sewer defect prediction can be conducted via EDL. Furthermore, the ambiguity of fused expert opinion $\mathcal{M}^F$ among multi-label nodes is mitigated via ESS as follows:

$$u_{\text{final}} = \text{ESS}(\mathcal{M}^F). \tag{4}$$

It should be noted that the ambiguity of $\mathcal{M}^F$ would only affect the result of uncertainty estimation and not $\mathbf{p}$.

B. Multi-segment Splitting Module

Based on the analysis of sewer videos, we can observe that the inspectors often manipulate the view of camera from far-to-near repeatedly. Such manipulation would naturally capture the various scales in terms of the global and local.Inspired by this observation, we introduce the optical flow to capture the moment of scale-variant. Specifically, given the frames of QV sewer video $\mathbf{X} = \{x_1, x_2, ..., x_T\}$, Farneback algorithm [42] is first utilized to estimate the optical flow based on two frames. Then, we cast the optical flow information in the four quadrants of the coordinate system. Since the text redaction of identification information would affect the incorrect estimation of optical flow, the optical flow information in the region of quadrant II is discarded. Finally, the global-scale segment X^G and local-scale segment X^L are captured based on the optical flow information (angle and magnitude) and temporal relation among frames, where the details of procedure is illustrated as the pseudocode in Alg. 1.

C. DST-based Combination Rule and Training of EDL

Based on subjective logic (SL) [41], EDL [24] is formulated to explicitly associate belief and uncertainty with Dirichlet distribution, forming the classification and uncertainty estimation in a unified framework. The multi-label classification (with K-dimension label $\mathbf{y} \in \mathbf{R}^{K\times 1}$) would be cast as multiple binary classifications (with K binary label $\hat{\mathbf{y}} \in \mathbf{R}^{K\times 2}$), since EDL could not be directly conducted for the multi-label classification. To simplify the formulation, an evidential deep learning network E is designed, representing the multi-scale segments as evidences $\mathbf{e} \in \mathbf{R}^{K\times 2}$. Then, for instance of X^G, the expert opinion $\mathcal{M}^G$ can be derived following [24]:

$$\mathcal{M}^G = \{\mathbf{b}^G, \mathbf{u}^G, \boldsymbol{a}^G\}, \tag{5}$$

where $\mathbf{b}^G = \{(b^G_{k,+}, b^G_{k,-})\}^K_{k=1}$, $\mathbf{u}^G = \{u^G_k\}^K_{k=1}$, and $\boldsymbol{a}^G = \{(a^G_{k,+}, a^G_{k,-})\}^K_{k=1}$ are the belief, uncertainty, and base rate of X^G, respectively. $+$ and $-$ denote the defective and non-defective nodes of the k-th defect category, respectively.

To explore the multi-scale complementarity among segments, we introduce the DST-based combination rule to fuse the $\mathcal{M}^G$ and $\mathcal{M}^L$, which can be formulated as follows:

Algorithm 1 Pseudocode of MSM in a Python-like style.

```
import math
def scale_variant_moment_capturing(X):
  XG_moment,XL_moment,X_others=[],[],[]
  for t,x_t in enumerate(X[:-1]):
    #ori_t and mag_t are the angle and magnitude vectors
    ori_t,mag_t = Farneback(x_t,x_tplus1)
    #C_ld,C_rd,C_ru,C_lu denote the direction sets:
    #↙, ↘, ↗, and ↖
    C_ld,C_rd,C_ru,C_lu=[],[],[],[]
    for ori_t(h,w),mag_t(h,w) in enumerate(ori_t,mag_t):
      #xi_mag is a threshold ξ_mag
      if mag_t(h,w)<xi_mag or (h,w) in quadrant_II:continue
      #Statistics of the four directions
      if -math.pi/2<=ori_t(h,w)<=0:
        C_ld.append((h,w))
      elif -math.pi<=ori_t(h,w)<-math.pi/2:
        C_rd.append((h,w))
      elif 0<ori_t(h,w)<=math.pi/2:
        C_ru.append((h,w))
      elif math.pi/2<ori_t(h,w)<=math.pi:
        C_lu.append((h,w))
    #Capturing the moment of scale-variant
    #I,III,IV denote the three quadrants
    #C_ld_I presents the entity (h,w) of C_ld in I quadrant
    #xi_mom is a threshold ξ_mom
    if (len(C_ld_I)-len(C_ru_I))>xi_mom and \
      (len(C_ru_III)-len(C_ld_III))>xi_mom and \
      (len(C_lu_IV)-len(C_rd_IV))>xi_mom:
      XG_moment.append(x_t)
    elif (len(C_ru_I)-len(C_ld_I))>xi_mom and \
      (len(C_ld_III)-len(C_ru_III))>xi_mom and \
      (len(C_rd_IV)-len(C_lu_IV))>xi_mom:
      XL_moment.append(x_t)
    else:
      X_others.append(x_t)
  return XG_moment,XL_moment,X_others

def temporal_relation_analysis(X,XG_moment,XL_moment,X_others):
  XG,XL=[],[]
  #x.indexOf(X) returns a time stamp of x in X.
  XG_moment_idx = [x.indexOf(X) for x in XG_moment]
  XL_moment_idx = [x.indexOf(X) for x in XL_moment]
  for x_others in X_others:
    x_others_idx=x_others.indexOf(X)
    if x_others_idx==0:
      XG.append(X_others)
    elif x_others_idx in XL_moment_idx:
      XL.append(X_others)
    elif x_others_idx in XG_moment_idx:
      XG.append(X_others)
  return XG,XL
#if __name__=='__main__':
X=from_QV_Sewer_Inspection()
XG_moment,XL_moment,X_others=scale_variant_moment_capturing(X)
XG,XL=temporal_relation_analysis(X,XG_moment,XL_moment,X_others)
```

$$\mathcal{M}^F = \mathcal{M}^G \oplus \mathcal{M}^L. \tag{6}$$

Specifically, the combination of the k-th $+$ node in terms of belief, uncertainty, and base rate can be formulated as follows:

$$\begin{cases} b^F_{k,+} = \dfrac{\text{Har}_+}{(1-\text{Con}_+)}, u^F_k = \dfrac{u^G_k u^L_k}{(1-\text{Con}_+)} \\ a^F_{k,+} = \dfrac{a^G_{k,+}(1-u^G_k) + a^L_{k,+}(1-u^L_k)}{2-u^G_k-u^L_k} \quad \text{for } u^G_k + u^L_k < 2\,, \\ a^F_{k,+} = \dfrac{a^G_{k,+} + a^L_{k,+}}{2} \qquad\qquad\qquad\qquad \text{for } u^G_k = u^L_k = 1 \end{cases} \tag{7}$$

where $\text{Har}_+ = b^G_{k,+}b^L_{k,+} + b^G_{k,+}u^L_k + b^L_{k,+}u^G_k$ denotes the relative harmony between opinions, $\text{Con}_+ = \sum_{k_1\neq k_2} b^G_{k_1,+}b^L_{k_2,+}$ presents the relative conflict between opinions. The k-th $+$ node can be formulated via 7 intuitively. Then, the prediction probability $\mathbf{p}$ of K binary classifications can be induced based on $\mathcal{M}^F$ via EDL.

Following the training stage of [24], we introduce the Type II Maximum Likelihood $\mathcal{L}_{EDL}$ and Kullback-Leibler (KL) divergence regularization $\mathcal{L}_{KL}$ to optimize the θ_E parameters of E. Here, the segment-wise and fusion-wise losses are both formulated as follows:

$$\mathcal{L}(\theta_E)_{overall} = \sum_{s\in\{G,L\}} \lambda_s(\mathcal{L}^s_{EDL} + \mathcal{L}^s_{KL}) + \lambda_F(\mathcal{L}^F_{EDL} + \mathcal{L}^F_{KL}), \tag{8}$$

where λ_s and λ_F are set to 0.5 and 1, respectively.

D. Evidence Selecting Scheme

Since the multi-label classification is cast as multiple binary classifications,we should design a method to estimate the final uncertainty score u_{final} among K fused expert opinions. In [19], the uncertainty score via aggregation operation, such as Max, Top-5, and Sum, can be concluded as the isolated-wise method, where the uncertainty of expert opinion is regarded as the uncertainty estimation intuitively. The belief of expert opinion is ignored in the isolated-wise method, resulting in the ambiguity of uncertainty estimation.

Here, we further analyze such ambiguity based on the procedure of isolated-based method. For instance, there is a simplified expert opinion $\{(b_{k,+}, b_{k,-}, u_k)\}_{k=1}^K$, where $b_{k,+}$, $b_{k,-}$, and u_k satisfy the constraint $b_{k,+} + b_{k,-} + u_k = 1$. Then, an uncertainty set $\mathbf{U} = \{u_{I_k}\}_{k=1}^K$ can be constructed, where I_k is the index of the k-th top value in $\mathbf{U}$. The isolated-based method estimates the uncertainty score can be formulated as:

$$u_{\text{Max}} = u_{I_1}, u_{\text{Top-5}} = \sum\nolimits_{k=1}^{5} u_{I_k}, u_{\text{Sum}} = \sum\nolimits_{k=1}^{K} u_{I_k}, \tag{9}$$

where u_{Max}, $u_{\text{Top-5}}$, and u_{Sum} are the uncertainty scores based on Max, Sum, and Top-5, respectively. Intuitively, the higher magnitude of u_{I_k} means that the query sample is in-distribution (ID), while the lower magnitude of u_{I_k} means that the query sample is OOD. Nevertheless, the isolated-based method is vulnerable, since Ambiguous Uncertainty (AU) cases of expert opinion would disturb the result of uncertainty estimation. As shown in Fig. 2, Certain Uncertainty (CU) cases present the high magnitudes in terms of belief $(b_{k,+}, b_{k,-})$ and uncertainty (u_k), whose high discriminability of uncertainty is beneficial for the isolated-based method to estimate the uncertainty precisely. However, AU cases could not be ignored, which present a relatively low uncertainty for OOD query sample, and a relatively high uncertainty for ID query sample, resulting in the ambiguous uncertainty for isolated-based method.

To mitigate such ambiguity, we propose a simple and efficient method termed as evidence selecting scheme (ESS), which introduces $b_{k,+}$ and $b_{k,-}$ in the procedure of uncertainty estimation to filter the AU case as follows:

$$\hat{\mathbf{U}} = \mathbf{U} \backslash \{u_{I_k} | \delta(b_{k,+}, b_{k,-}) + u_{I_k} < \xi_{ESS}\}_{k=1}^K, \tag{10}$$

where $\delta(b_{k,+}, b_{k,-}) = |b_{k,+} - b_{k,-}|$, and ξ_{ESS} is set to 0.6. If $\hat{\mathbf{U}} = \emptyset$, we assign u_{I_1} as a single entity to $\hat{\mathbf{U}}$. Based on the filtered $\hat{\mathbf{U}}$, the disambiguated uncertainty score u_{final} is obtained via Eq. 9.

IV. Experiments

A. Experimental Setup

1) Dataset: VideoPipe [27] is a large-scale benchmark dataset collected from real-world sewer pipes, consisting of two sewer pipe defect datasets: QV-Pipe and CCTV-Pipe based on the different acquisition devices. Here, we focus on the sewer defect detection in the QV sewer videos, while QV-Pipe is utilized to conduct the evaluation experiments of SEFN. QV-Pipe consists of 9,601 QV sewer videos annotated with 17-dimension multi-label (1 normal category and 16 sewer defect categories), where the total duration of videos exceeds 55 hours with about 4,950,000 frames. Since the annotations of testing set are not public, we conduct the experiments focused on the training and validation sets.

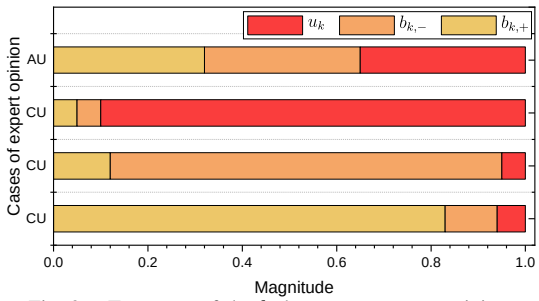


Fig. 2. Toy cases of the k-th category expert opinion.

2) Implementation Details: SEFN is implemented with MMAction2 framework [45], while the experiments are conducted on NVIDIA Tesla A100 GPUs. Video swin transformer [46] is utilized as the backbone for the evidential deep learning network E whose parameters are initialized from Kinetics-600 (K-600) pre-trained models [47]. In order to evaluate SEFN in terms of validity and reliability, multi-label sewer defect classification $\mathcal{T}_{cls}$ and OOD detection $\mathcal{T}_{ood}$ are introduced, where the details are as follows: (1) For $\mathcal{T}_{cls}$, video swin transformer [46] is utilized as the backbone for the feature extractor F whose parameters are initialized from Kinetics-600 (K-600) pre-trained models [47]. ξ_{mag} and ξ_{mom} are set to 50 and 5, respectively. AdamW optimizer [48] with a weight decay of $1e-3$ is adopted to train the network. The initial learning rate is set to $1e-4$ and the total epochs are 100. For each segment, we sample 12 frames with a temporal stride 8 at a random start point. The frames are resized to 224×224, and the batch size is set to 8 for all experiments. (2) For $\mathcal{T}_{ood}$, the details of network structure and training are the same as $\mathcal{T}_{cls}$. Then, we design two settings to define the unknown samples, termed as Most-4 and Least-4, respectively. The first setting is to select the samples with the four most numerous categories (CJ, CK, ZW and JG) as the unknown samples, while the second setting defines the unknown samples based on the samples with the four least numerous categories (QF, SL, TL and CQ).

3) Evaluation Metrics: Following [17], [19], we adopt the five metrics to evaluate the performances in terms of validity and reliability, including Mean Average Precision (**mAP**), Mean F1 (**mF1**) score, **AUROC**, **AUPR**, and **FPR95**.

B. Comparisons With Other Methods

1) Multi-label Sewer Defect Classification $\mathcal{T}_{cls}$: Since there are only a few works on QV-Pipe, we reproduce the state-of-art methods in the community of video classification, which can be categorized in terms of different backbones (I3D [49], Resnet (Res-50) [55], and video swin transformer (Swin-B) [46]). Moreover, Super-Event [32], PDAN [33], and MS-

TABLE I

COMPARISON WITH THE STATE-OF-THE-ART METHODS. MAP(%)↑, MF1(%)↑, AND $F1_{NORMAL}$(%)↑ ARE UTILIZED AS THE METRICS FOR $\mathcal{T}_{cls}$. THE BEST RESULTS ARE **BOLD**.

Methods	Backbone	Pretrain	mAP	mF1	$F1_{Normal}$
I3D [49]	I3D	K-400	17.24	21.87	40.93
Super-Event [32]	I3D	K-400	34.86	31.44	39.80
PDAN [33]	I3D	K-400	37.96	43.39	58.77
MS-TCT [34]	I3D	K-400	39.40	44.46	66.03
SlowFast [50]	Res-50	K-400	44.89	54.68	78.84
TIN [51]	Res-50	K-400	46.57	55.31	74.78
TANet [52]	Res-50	K-400	48.23	31.94	45.49
TSM [53]	Res-50	K-400	53.38	49.65	70.11
SVST [54]	Swin-B	K-400	57.30	49.44	75.49
SEFN (Ours)	Swin-B	K-400	60.66	66.13	87.69
SVST [54]	Swin-B	K-600	58.56	55.97	78.26
SEFN (Ours)	Swin-B	K-600	**61.47**	**67.99**	**89.20**

TABLE II

COMPARISON WITH THE STATE-OF-THE-ART OOD METHODS. AUROC(%)↑, AUPR(%)↑, AND FPR95(%)↓ ARE UTILIZED AS THE METRICS FOR $\mathcal{T}_{ood}$. THE BEST RESULTS ARE **BOLD**.

Methods	$\mathcal{T}_{ood}$ Least-4	$\mathcal{T}_{ood}$ Most-4
	AUROC/AUPR/FPR95	AUROC/AUPR/FPR95
MaxLogit [57]	71.14/96.63/84.94	73.67/75.60/85.19
MaxProb [58]	53.16/92.57/89.73	54.17/46.08/84.88
JointEnergy [58]	70.70/96.56/86.30	72.93/74.14/84.88
MSP [35]	69.13/96.74/87.67	74,85/80.22/84,67
SLCS [59]	71.90/96.73/85.62	73.55/75.07/83.23
MaxCosine [60]	71.46/96.54/82.19	73.05/71.97/83.02
MaxNorm [60]	61.89/95.66/89.73	74.21/80.45/89.40
DML [60]	71.14/96.63/55.62	77.91/83.11/54.77
SEFN w/o ESS	72.79/87.82/67.40	78.37/70.32/43.86
SEFN (Ours)	**84.91/97.83/45.39**	**80.17/83.76/43.57**

TCT [34] are designed based on the deep features extracted via I3D [49] pretrained on Kinetics-400 (K-400) [56], while the backbones of other end-to-end methods (I3D [49], SlowFast [50], TIN [51], TANet [52], TSM [53], SVST [27]) are pretrained on K-400/600. In order to investigate the effect of backbones pretrained on different dataset, we design a version of SEFN whose backbone is pretrained on K-400. As reported in Tab. I, we can observe that SEFN is superior to the other state-of-art methods. The second best method is SVST [54], which is a video swin transformer-based multi-label sewer defect classification method. In contrast, SEFN with backbone pretrained on K-400/600 outperforms SVST by 3.36%/2.91%, 16.69%/12.02%, and 12.20%/10.94% in terms of mAP, mF1, and $F1_{Normal}$, respectively. The main difference between SEFN and SVST is that SVST rigidly represents the sewer defect based on the completed duration segment, resulting in ignoring the complementarity among the different scale segments. Such results argue the superiority of SEFN, benefiting from the multi-scale complementarity to precisely generate the decision-making, especially for judging the defects ($F1_{Normal}$ = **89.20%**). Meanwhile, the methods with transformer-based backbone are superior to I3D-based and Resnet-based comprehensively, revealing the strong capability of transformer for representing the sewer defect.

2) OOD Detection $\mathcal{T}_{ood}$: We compare our method against the competitive OOD detection methods of multi-label classification, including the probability-based methods (MaxProb [58]), logit-based methods (MaxLogit [57], JointEnergy [58], MSP [35], and SLCS [59]), and methods based on the last-layer features and weights (MaxCosine [60], MaxNorm [60], and DML [60]). Moreover, SEFN without ESS (SEFN w/o ESS) is conducted to evaluate the effect of ESS, where the summation is utilized to aggregate the uncertainties of all categories as OOD score. As reported in Tab. II, SEFN w/o ESS achieves the competitive results in terms of AUROC and FPR95. However, the improvement in terms of AUROC is not substantial, and we even observe a decrease in AUPR. In Section III-D, we provide a detailed analysis of the reason behind the ambiguous uncertainty estimation, in which AU would degenerate the performance of uncertainty estimation. It can be seen that for $\mathcal{T}_{ood}$ Least-4, SEFN surpasses the second best method DML 13.77%, 1.20%, and 10.23% in terms of AUROC, AUPR, and FPR95, respectively. For $\mathcal{T}_{ood}$ Most-4, SEFN outperforms DML 2.26%, 0,65%, and 11.20% in terms of AUROC, AUPR, and FPR95, respectively. These facts verify that ESS can significantly mitigate the interference of AU, enabling the trustworthy performance of SEFN on the uncertainty estimation of the samples from unknown category.

TABLE III

ABLATION ANALYSIS OF FSM ON QV-PIPE. SEFN† DONATES A VARIANT SEFN IS DESIGNED TO ADAPT THE SINGLE SEGMENT, WHERE DST-BASED COMBINATION IS DISCARDED IN SEFN†. THE BEST RESULTS ARE **BOLD**.

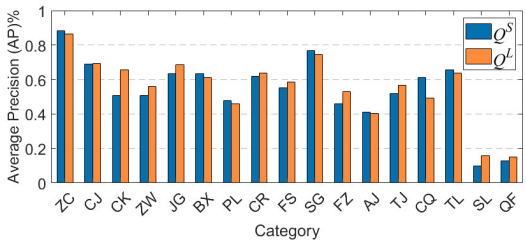
	w/ MSM	Segment	mAP	mF1	$F1_{Normal}$
SEFN†	✓	Q^L	58.70	65.25	86.16
	✓	Q^S	55.09	66.07	88.47
SEFN		Q^{B_1} & Q^{B_2}	58.19	64.37	86.84
	✓	Q^S & Q^L	**61.47**	**67.99**	**89.20**

C. Ablation Studies

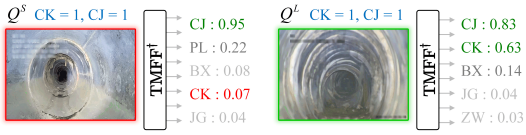
1) The Effect of Multi-scale Complementarity and MSM: To validate the effect of multi-scale complementarity, we conduct the ablation experiments as reported in Tab. III, which can be divided into two settings.

The first setting is to utilize only a single segment from Q^G or Q^L for our method. A variant SEFN (SEFN†) is designed to adapt the single segment, where DST-based combination is discarded in SEFN†. It can be seen that the performances of SEFN† fed with the single focal length segment are not superior to SEFN with Q^G and Q^L. Moreover, we further investigate the individual characteristics of Q^G and Q^L. As shown in Fig. 3(a), the performances of SEFN† for each category are reported separately. There is a fact that the performance of SEFN† for the major categories can benefit from the local information of Q^L (Fig. 3(b)). Meanwhile, the appearances of defects with global context, such as CK (Fig. 3(c)), can be captured reliably via the global information of Q^G. Thus, the information from Q^G and Q^L are both vital for representing the sewer defects, revealing the complementarity among the segments.

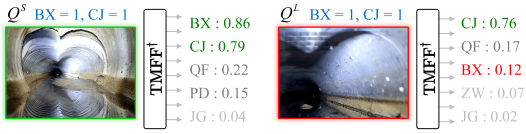
In the second setting, we aim to verify the effect of MSM. bisecting clipping (BC) is designed to replace MSM, which clips the video into two segments with the same length, termed



(a) Comparison of SEFN† fed with Q^G or Q^L.



(b) Classification results from SEFN† fed with Q^G and Q^L, respectively. The defect categories of sample are CK and CJ.



(c) Classification results from SEFN† fed with Q^G and Q^L, respectively. The defect categories of sample are BX and CJ.

Fig. 3. Quantitative and qualitative analysis for the performance of SEFN†.

TABLE IV
ABLATION ANALYSIS OF MSM AND DST-BASED COMBINATION. C1, C2, AND C3 DENOTE THREE VERSIONS OF SEFN CONDUCTED VIA DIFFERENT STRUCTURES. THE BEST RESULTS ARE **BOLD**.

	w/ MSM	w/ DST	$\mathcal{T}_{cls}$	$\mathcal{T}_{ood}$ Least-4	$\mathcal{T}_{ood}$ Most-4
			mAP/mF1/F1$_{\text{Normal}}$	AUROC/AUPR/FPR95	AUROC/AUPR/ FPR95
C1			58.56/55.97/78.26	69.15/89.51/68.82	77.19/69.15/43.78
C2	✓		60.99/63.77/80.10	71.82/88.45/66.63	78.52/71.40/44.68
C3	✓	✓	**61.47/67.99/89.20**	**84.91/97.83/45.39**	**80.17/83.76/43.57**

TABLE V
ABLATION ANALYSIS OF ESS. THE BEST RESULTS ARE **BOLD**.

	w/ ESS	$\mathcal{T}_{ood}$ Least-4	$\mathcal{T}_{ood}$ Most-4
		AUROC/AUPR/FPR95	AUROC/AUPR/FPR95
Max		66.90/89.76/68.77	76.30/67.10/43.88
Top-5		72.02/88.32/68.66	77.18/68.86/44.47
Sum		72.79/87.82/67.40	78.37/70.32/43.86
Max	✓	82.96/96.17/46.35	76.05/79.40/51.06
Top-5	✓	84.15/97.40/47.61	77.05/80.21/47.34
Sum	✓	**84.91/97.83/45.39**	**80.17/83.76/43.57**

as Q^{B_1} and Q^{B_2}. It can be observed that the performances of SEFN are degenerated based on the clipping operation BC. Such results further demonstrate that the multi-focus segments split via FSM present the complementary information among Q^L and Q^S, improving the performances of SEFN.

2) The Effect of MSM and DST-based combination: To clarify the effect of MSM and DST-based combination, we conduct the ablation experiments for $\mathcal{T}_{cls}$ and $\mathcal{T}_{ood}$, as reported in Tab. IV. Three versions of SEFN are conducted via the different settings of FSM and DST-based combination, termed as C1, C2, and C3, where the concatenating operation of features is utilized to replace DST-based combination in C1 and C2. It can be seen that MSM and DST-based combination show the incremental improvements for $\mathcal{T}_{cls}$ simultaneously.

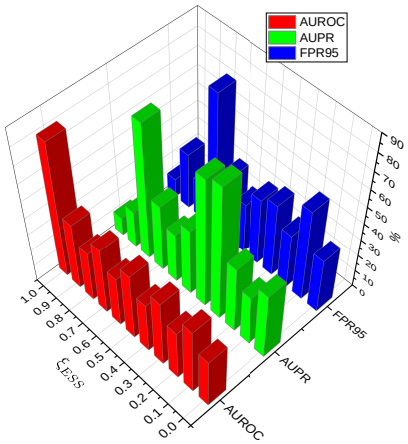


Fig. 4. Results of the sensitivity experiment for ξ_{ESS} on $\mathcal{T}_{ood}$ Most-4.

However, compared with the results of C1 and C2, the improvement of introducing MSM for $\mathcal{T}_{ood}$ is slight. We argue that the concatenating operation might not aggregate the multi-scale complementarity in terms of uncertainty. In the result of C3, the significant improvements, especially for $\mathcal{T}_{ood}$ Least-4, reveal the effect of DST-based combination to joint the expert opinions to estimate uncertainty.

3) The Effect of ESS: To demonstrate the effect of ESS, we compare the isolated-based methods with their ESS-based variants. As reported in Tab. V, EDS significantly improves the performances of isolated-based methods for $\mathcal{T}_{ood}$. These results demonstrate that the ambiguity caused by AU cases weakens the reliability of uncertainty estimation, verifying the effect of ESS to discard CC. Moreover, we investigate the sensitivity of ξ_{ESS}. As shown in Fig. 4, the results is reported on $\mathcal{T}_{ood}$ Most-4 with the various ξ_{ESS} from 0 to 1. It can be observed that the most reliable performance is achieved when $\xi_{ESS} = 0.6$. Meanwhile, the performances become worse with the increase of magnitude. The reason might be that the higher magnitude of ξ_{ESS} would reduce the "useful" CU cases.

4) The Limitation of SEFN: The main limitation of SEFN is the inference speed. The inference speeds of each module for a XXX second QV sewer video are reported as follows: XXX ms (MSM), XXX ms (EDL), and XXX ms (ESS). Such inference speed presents that SEFN is acceptable for the offline QV-based inspection system, while it might not be tolerant to the online system with the robot vehicle. Thus, we would explore an efficient splitting method for scale variant in future works.

V. CONCLUSION

In this paper, we propose a segment-aware evidential fusion network (SEFN) for multi-label sewer defect classification in quick view (QV)-based sewer video, in which the issues of capturing the multi-scale complementarity and disambiguating the uncertainty estimation are addressed. Experimental results demonstrate that SEFN achieves the superior performances in terms of multi-label sewer defect classification and uncertainty estimation. In future works, the improvement of SEFN would be further explored in terms of inference speed.

REFERENCES

[1] D.-H. Koo and S. T. Ariaratnam, "Innovative method for assessment of underground sewer pipe condition," *Autom. Constr.*, vol. 15, no. 4, pp. 479–488, 2006.

[2] J. Curiel-Esparza, J. Canto-Perello, and M. A. Calvo, "Establishing sustainable strategies in urban underground engineering," *Sci. Eng. Ethics.*, vol. 10, pp. 523–530, 2004.

[3] E. Kuliczkowska, "The interaction between road traffic safety and the condition of sewers laid under roads," *Transp. Res. D Transp. Environ.*, vol. 48, pp. 203–213, 2016.

[4] A. Kakogawa, Y. Komurasaki, and S. Ma, "Anisotropic shadow-based operation assistant for a pipeline-inspection robot using a single illuminator and camera," in *IROS*, 2017, pp. 1305–1310.

[5] D. Li, Q. Xie, Z. Yu, Q. Wu, J. Zhou, and J. Wang, "Sewer pipe defect detection via deep learning with local and global feature fusion," *Autom. Constr.*, vol. 129, p. 103823, 2021.

[6] L. M. Dang, S. I. Hassan, S. Im, I. Mehmood, and H. Moon, "Utilizing text recognition for the defects extraction in sewers cctv inspection videos," *Comput Ind.*, vol. 99, pp. 96–109, 2018.

[7] R. Rayhana, Y. Jiao, Z. Bahrami, Z. Liu, A. Wu, and X. Kong, "Valve detection for autonomous water pipeline inspection platform," *IEEE/ASME Trans. Mechatronics*, vol. 27, no. 2, pp. 1070–1080, 2022.

[8] R. Ren, T. Hung, and K. C. Tan, "A generic deep-learning-based approach for automated surface inspection," *IEEE Trans. Cybern.*, vol. 48, no. 3, pp. 929–940, 2018.

[9] C. Hu and Y. Wang, "An efficient convolutional neural network model based on object-level attention mechanism for casting defect detection on radiography images," *IEEE Trans. Ind. Electron.*, vol. 67, no. 12, pp. 10 922–10 930, 2020.

[10] P. Tan, X. Li, Z. Wu, J. Ding, J. Ma, Y. Chen, Y. Fang, and Y. Ning, "Multialgorithm fusion image processing for high speed railway dropper failure–defect detection," *IEEE Trans. Systems, Man, and Cybernetics: Systems*, vol. 51, no. 7, pp. 4466–4478, 2021.

[11] X. Tao, D. Zhang, Z. Wang, X. Liu, H. Zhang, and D. Xu, "Detection of power line insulator defects using aerial images analyzed with convolutional neural networks," *IEEE Trans. Systems, Man, and Cybernetics: Systems*, vol. 50, no. 4, pp. 1486–1498, 2020.

[12] Z. Chen, Y. Liao, J. Li, R. Huang, L. Xu, G. Jin, and W. Li, "A multi-source weighted deep transfer network for open-set fault diagnosis of rotary machinery," *IEEE Trans. Cybern.*, vol. 53, no. 3, pp. 1982–1993, 2023.

[13] Z. Zhang and L. Zhang, "Unsupervised pixel-level detection of rail surface defects using multistep domain adaptation," *IEEE Trans. Systems, Man, and Cybernetics: Systems*, pp. 1–12, 2023.

[14] Y. Wang, J. Wang, Y. Cao, S. Li, and O. Kwan, "Integrated inspection on pcb manufacturing in cyber–physical–social systems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 4, pp. 2098–2106, 2023.

[15] C. Piciarelli, D. Avola, D. Pannone, and G. L. Foresti, "A vision-based system for internal pipeline inspection," *IEEE Trans. Ind. Informat.*, vol. 15, no. 6, pp. 3289–3299, 2019.

[16] Q. Xie, D. Li, J. Xu, Z. Yu, and J. Wang, "Automatic detection and classification of sewer defects via hierarchical deep learning," *IEEE Trans. Autom. Sci. Eng.*, vol. 16, no. 4, pp. 1836–1847, 2019.

[17] J. B. Haurum and T. B. Moeslund, "Sewer-ml: A multi-label sewer defect classification dataset and benchmark," in *Proceedings of CVPR*, 2021, pp. 13 456–13 467.

[18] L. M. Dang, H. Wang, Y. Li, T. N. Nguyen, and H. Moon, "Defecttr: End-to-end defect detection for sewage networks using a transformer," *Constr. Build. Mater.*, vol. 325, p. 126584, 2022.

[19] C. Zhao, C. Hu, H. Shao, Z. Wang, and Y. Wang, "Towards trustworthy multi-label sewer defect classification via evidential deep learning," in *ICASSP*. IEEE, 2023, pp. 1–5.

[20] S.-J. Wang, Y. He, J. Li, and X. Fu, "Mesnet: A convolutional neural network for spotting multi-scale micro-expression intervals in long videos," *IEEE Trans. Image Process.*, vol. 30, pp. 3956–3969, 2021.

[21] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 20, no. 2, pp. 189–201, 2009.

[22] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[23] B. Roghani, F. Cherqui, M. Ahmadi, P. Le Gauffre, and M. Tabesh, "Dealing with uncertainty in sewer condition assessment: Impact on inspection programs," *Autom. Constr.*, vol. 103, pp. 117–126, 2019.

[24] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," vol. 31, 2018.

[25] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," vol. 31, 2018.

[26] G. Shafer, *A mathematical theory of evidence*. Princeton university press, 1976, vol. 42.

[27] Y. Liu, X. Zhang, Y. Li, G. Liang, Y. Jiang, L. Qiu, H. Tang, F. Xie, W. Yao, Y. Dai *et al.*, "Videopipe 2022 challenge: Real-world video understanding for urban pipe inspection," in *2022 26th ICPR*. IEEE, 2022, pp. 4967–4973.

[28] C. Chen, X. Zou, Z. Zeng, Z. Cheng, L. Zhang, and S. C. H. Hoi, "Exploring structural knowledge for automated visual inspection of moving trains," *IEEE Trans. Cybern.*, vol. 52, no. 2, pp. 1233–1246, 2022.

[29] P. Afshar, A. Mohammadi, K. N. Plataniotis, A. Oikonomou, and H. Benali, "From handcrafted to deep-learning-based cancer radiomics: Challenges and opportunities," *IEEE Signal Process. Mag.*, vol. 36, no. 4, pp. 132–160, 2019.

[30] C. Hu, B. Dong, H. Shao, J. Zhang, and Y. Wang, "Toward purifying defect feature for multilabel sewer defect classification," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–11, 2023.

[31] Y. Zhu, X. Li, C. Liu, M. Zolfaghari, Y. Xiong, C. Wu, Z. Zhang, J. Tighe, R. Manmatha, and M. Li, "A comprehensive study of deep video action recognition," *arXiv preprint arXiv:2012.06567*, 2020.

[32] A. Piergiovanni and M. S. Ryoo, "Learning latent super-events to detect multiple activities in videos," in *Proceedings of CVPR*, 2018, pp. 5304–5313.

[33] R. Dai, S. Das, L. Minciullo, L. Garattoni, G. Francesca, and F. Bremond, "Pdan: Pyramid dilated attention network for action detection," in *Proceedings of WACV*, 2021, pp. 2970–2979.

[34] R. Dai, S. Das, K. Kahatapitiya, M. S. Ryoo, and F. Brémond, "Ms-tct: multi-scale temporal convtransformer for action detection," in *Proceedings of CVPR*, 2022, pp. 20 041–20 051.

[35] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.

[36] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *ICML*. PMLR, 2015, pp. 1613–1622.

[37] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *ICML*. PMLR, 2016, pp. 1050–1059.

[38] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," vol. 30, 2017.

[39] J. Van Amersfoort, L. Smith, Y. W. Teh, and Y. Gal, "Uncertainty estimation using a single deep deterministic neural network," in *ICML*. PMLR, 2020, pp. 9690–9700.

[40] A. Damianou and N. D. Lawrence, "Deep gaussian processes," in *Artificial intelligence and statistics*. PMLR, 2013, pp. 207–215.

[41] A. Jøsang, *Subjective logic*. Springer, 2016, vol. 4.

[42] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scandinavian conference on image analysis*. Springer, 2003, pp. 363–370.

[43] T. Jamil and C. J. ter Braak, "Selection properties of type ii maximum likelihood (empirical bayes) in linear models with individual variance components for predictors," *Pattern Recognit. Lett.*, vol. 33, no. 9, pp. 1205–1212, 2012.

[44] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *ICLR*, 2017.

[45] M. Contributors, "Openmmlab's next generation video understanding toolbox and benchmark," https://github.com/open-mmlab/mmaction2, 2020.

[46] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of CVPR*, 2022, pp. 3202–3211.

[47] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A short note about kinetics-600," *arXiv preprint arXiv:1808.01340*, 2018.

[48] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[49] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of CVPR*, 2017, pp. 6299–6308.
[50] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of ICCV*, 2019, pp. 6202–6211.
[51] H. Shao, S. Qian, and Y. Liu, "Temporal interlacing network," in *Proceedings of AAAI*, vol. 34, no. 07, 2020, pp. 11 966–11 973.
[52] Z. Liu, L. Wang, W. Wu, C. Qian, and T. Lu, "Tam: Temporal adaptive module for video recognition," in *Proceedings of ICCV*, 2021, pp. 13 708–13 718.
[53] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *Proceedings of ICCV*, 2019, pp. 7083–7093.
[54] X. Fang, "2nd place winners: Icpr videopipe challenge - track on video defect classification," 2022. [Online]. Available: https://videopipe.github.io/results/track1-top2.pdf
[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of CVPR*, June 2016.
[56] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.
[57] D. Hendrycks, S. Basart, M. Mazeika, M. Mostajabi, J. Steinhardt, and D. Song, "A benchmark for anomaly segmentation," *arXiv preprint arXiv:1911.11132*, vol. 1, no. 2, p. 5, 2019.
[58] H. Wang, W. Liu, A. Bocchieri, and Y. Li, "Can multi-label classification networks know what they don't know?" vol. 34, 2021, pp. 29 074–29 087.
[59] L. Wang, S. Huang, L. Huangfu, B. Liu, and X. Zhang, "Multi-label out-of-distribution detection via exploiting sparsity and co-occurrence of labels," *Image Vis. Comput.*, vol. 126, p. 104548, 2022.
[60] Z. Zhang and X. Xiang, "Decoupling maxlogit for out-of-distribution detection," in *Proceedings of CVPR*, 2023, pp. 3388–3397.
[61] P. Xu, L. Xiao, B. Liu, S. Lu, L. Jing, and J. Yu, "Label-specific feature augmentation for long-tailed multi-label text classification," in *Proceedings of AAAI*, vol. 37, no. 9, 2023, pp. 10 602–10 610.
[62] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, 2021.