IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 10 January 2024, accepted 12 March 2024, date of publication 16 April 2024, date of current version 26 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3390029

RESEARCH ARTICLE

SYRFA: SYnthetic-to-Real Adaptation via Feature Alignment for Video Anomaly Detection

JONGHWAN HONG, (Student Member, IEEE), BOKYEUNG LEE, KYUNGDEUK KO, (Graduate Student Member, IEEE), AND HANSEOK KO, (Senior Member, IEEE)
School of Electrical Engineering, Korea University, Seoul 02841, South Korea

Corresponding author: Hanseok Ko (hsko@korea.ac.kr)

This work was supported by the "Development of Cognitive/Response Advancement Technology for AI Avatar Commercialization" project funded by the Brand Engagement Network (BEN) under Grant Q2312881.

ABSTRACT Video Anomaly Detection (VAD) has garnered significant attention in computer vision, especially with the exponential growth of surveillance videos. Recently, the synthetic dataset has been released to address the imbalance problem between normal and abnormal scenarios in real-world datasets by providing various combinations of events. Motivated by the release of synthetic datasets, many studies have attempted to handle domain shifts by generating synthetic-real or real-synthetic abnormal scenarios. However, these approaches still suffer from a substantial computation burden due to the generation model. In this paper, we aim to alleviate the domain gap without relying on any generation model. We propose a novel framework named the SYnthetic-to-Real via Feature Alignment (SYRFA) for VAD. The SYRFA consists of two learning phases: learning synthetic knowledge and adaptation to the real-world domain. These two learning phases facilitate the incorporation of rich synthetic knowledge into the real-world domain. To address the domain shift between synthetic and real domains, we introduce consistency learning, aligning feature representations to map closely between the synthetic and real-world domains. Additionally, in the adaptation phase, we propose the Residual Additional Parameters (RAP), a simple yet effective approach for handling domain gaps. RAP is designed with a residual path for learning local patterns, crucial in VAD due to circumstantial feature representation. It contributes to obtaining transferable feature representations with fewer additional computations. The proposed framework demonstrates superior performance on VAD benchmark datasets. Especially, Our framework outperforms other methods by a margin of 0.8% on ShanghaiTech. Moreover, the ablation study highlights the effectiveness of the proposed framework and RAP.

INDEX TERMS Video anomaly detection, domain adaptation, synthetic-to-real.

I. INTRODUCTION

Detection plays a pivotal role in the realm of conventional surveillance tasks in computer vision [1], [2], [3], [4]. Within the domain of surveillance and factory automation, Video Anomaly Detection (VAD) has emerged as a critical pursuit. This significance is attributed to the exponential proliferation of surveillance video data, underlining the pressing need for effective anomaly detection mechanisms to enhance security and operational efficiency. VAD aims to identify exceptional events that deviate from typical scenarios. In a typical surveillance scenario, VAD operates under an open-set condition, characterized by an unbounded categorization of normal and abnormal samples. The challenges in VAD arise from the existence of such open-set conditions and the scarcity of training data related to abnormal events. Accordingly, the supervised learning approach encounters inherent limitations due to the impracticality of achieving a balanced training dataset encompassing both normal and abnormal scenarios. Therefore, most studies treat the VAD task as a one-class

The associate editor coordinating the review of this manuscript and approving it for publication was Zijian Zhang.

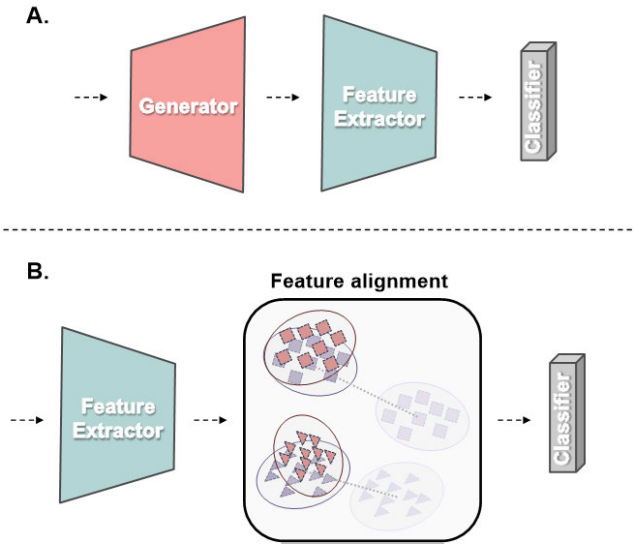

FIGURE 1. A comparison with existing methods and ours. Existing methods using synthetic data generate the synthetic-real or real-synthetic scenarios with generation model (A). Our method can render the transferable feature representation via feature alignments without any generation model (B).

classification problem, training exclusively on normal events while subjecting both normal and abnormal events to testing. With this paradigm, methods for VAD are categorized into three approaches: reconstruction-based methods, prediction-based methods, and synthetic-to-real-based methods.

Reconstruction-based methods are designed to construct networks capable of restoring input frames, thereby flagging anomalies through the identification of samples associated with high reconstruction errors. In contrast, prediction-based methods formulate models to generate missing frames in pursuit of temporal consistency, subsequently discerning anomalies as deviations between predicted frames and corresponding ground truth. While Such approaches show significant performance, examples with high reconstruction or prediction error do not contribute to adequate solutions for detecting anomalies due to the implicit generalization ability of deep networks. Such robust deep networks yield well-reconstructed or predicted samples.

To mitigate the dependence solely on the knowledge related to normal distribution, Acsintoae et al. [5] propose the synthetic VAD dataset, denoted as UBnormal. This dataset encompasses diverse scenarios crafted by 3D animators and 2D backgrounds. Motivated by the UBnormal dataset, Acsintoae et al. [5] and Liu et al. [6] proposed an innovative framework for rendering real-to-synthetic and synthetic-to-real abnormal samples, addressing the challenges of unbounded categories and the scarcity of abnormal training data, as shown in Fig. 1 (A). Although these generation methods for VAD demonstrate notable improvements in public VAD datasets, they neglect to handle some challenges. First, These methods primarily focus on an explicit approach to generating abnormal samples and do not consider the implicit domain shift between real and synthetic data. It is crucial to learn domain-invariant representations to alleviate domain shifts. Second, their methodologies demand significant computational resources for generating novel abnormal samples. These samples are derived from auxiliary generation networks, such as Cycle-GAN and VAE, necessitating the training of the feature extractor using both synthetic and real data. This training workflow inevitably leads to an increase in the number of network parameters, memory consumption, and training time. Third, the application of these methods to unseen data domains necessitates pre-training of the additional networks, as their effectiveness relies on the generation network during the training process. Therefore, the resource-intensive training process becomes indispensable when striving to generate high-quality data for learning from unseen domains.

Our method is motivated by such challenges and we propose a novel framework called SYnthetic-to-Real adaptation via Feature Alignment (SYRFA) to solve the problems, as shown in Fig. 1 (B). The proposed SYRFA operates in two distinct phases: ''Learning on Synthetic Knowledge'' and ''Adaptation to the Real Domain'', as depicted in Fig. 2. In the ''Learning on Synthetic Knowledge'' phase, the feature extractor is trained on synthetic data containing the various combinations of scene and action categories to provide abundant feature representation. In the 'Adaptation to the Real Domain' phase, the feature extractor adapts to the real domain via feature alignment, aiming to minimize the difference among source and target domains for learning domain-invariant representations. In contrast to previous works such as [5] and [6], our framework leverages the extensive knowledge provided by synthetic data to enhance its capability in the real data domain. This is achieved through learning domain-invariant representations without the reliance on any generation models, thus allowing for the design of an efficient network architecture. In addition, inspired by [7] which applies the adaptive parameters for adaptation, we introduce the Residual Additional Parameters (RAP). RAP is a developed version of adaptive parameters, proving more suitable for the VAD task due to its capacity to leverage local patterns with minimal computational burden. RAP effectively addresses domain shifts by aligning different domain features, leading to improved performance.

The main contributions are summarized as follows:

- We propose a novel framework SYRFA which leverages the plentiful knowledge of synthetic data by learning the domain-invariant representation without reliance on any generation.
- We propose the Residual Additional Parameters (RAP) which is effective for the VAD task and hase fewer computational burdens. The RAP shows effective adaptation abilities from the synthetic domain to the real domain.
- Our SYRFA demonstrates the effectiveness of the proposed framework on VAD benchmark datasets. In the ablation study, extensive experiments show that

the proposed framework and RAP perform significant results without any generation model.

II. RELATED WORK

A. VIDEO ANOMALY DETECTION

The purpose of VAD is to detect abnormal events in videos. Traditional VAD methods [8], [9], [10], [11] mainly rely on hand-crafted features or classical machine learning techniques. For instance, Adam et al. [8] describe the normal local histogram of optical low-level observations. Cong et al. [10] introduce a sparse reconstruction cost to measure normality. However, these methods face challenges in handling complex scenarios. Recently, deep learning methods [6], [12], [13], [14], [15], [16], [17], [18] have demonstrated superior performance, leveraging the powerful representation abilities of neural networks. Many studies, including reconstruction and prediction methods, train on the normal distribution and aim to detect out-of-distribution events, considering the lack of abnormal scenarios in real datasets. Reconstruction-based methods [15], [19], [20] focus on restoring the input frames and detecting the abnormalities through high reconstruction errors. Hasan et al. [19], for example, use an autoencoder to extract the features and calculate the scores based on the reconstruction errors. In contrast, prediction methods [14], [16], [21] emphasize temporal consistency. These methods learn to predict missing frames and identify the anomalies through the difference between the ground truth and the predicted frame. However, due to the powerful generalization capabilities of deep neural networks, their outputs are well generated and challenging to distinguish from the corresponding ground truth. To address the limitations of poor abnormal scenarios in real datasets, Acsintoae et al. [5] release the synthetic dataset. Studies on synthetic data propose novel approaches for generating real-to-synthetic and synthetic-to-real samples. However, these approaches have two limitations as they do not consider domain-invariant representation to reduce domain differences and require additional generation models such as VAE and Cycle-GAN. In this work, we present a novel framework to alleviate these issues.

III. PROPOSED METHODS

Our method is divided and organized into (1) learning the synthetic knowledge and (2) adaptation to the real domain. The feature extractor is trained on synthetic data for learning the various combinations of scene and action categories in advance (learning on synthetic knowledge). Subsequently, our network is adapted to the real domain using a self-supervised manner for reducing the synthetic-real domain gap (adaptation to real domain). To address the domain gap, we leverage the proposed RAP which can obtain transferable feature representation based on synthetic knowledge.

A. LEARNING ON SYNTHETIC KNOWLEDGE

In contrast to real datasets, which suffer from imbalanced information distribution between normal and abnormal scenarios, synthetic datasets offer rich diverse scenarios. Therefore, the goal of this section is to present various scenarios using synthetic data that mimic real-world occurrences. As shown in Fig. 2, the propagation of learning on synthetic knowledge is indicated as a red arrow path. The input consists of synthetic data obtained from object-level frames using an object detector, represented as $X^s \in \mathbb{R}^{C\times T\times H\times W}$ where C, T, H, W represent channel size, temporal length, height, and width, respectively. The normal and abnormal frames within the synthetic data are denoted as X^s_{nor} and X^s_{abn}, respectively. The feature extractor network and n-th block of feature extractor denoted by $f_\theta(\cdot)$ and $f^n_\theta(\cdot)$, respectively. The classifier for classification loss is represented as f_ϕ. The feature extractor comprises three blocks and yields two types of representations: the feature representation of the input and the augmented feature representation via the augmentation technique from [22]. It can be formulated as

$$O_1, O'_1 = f^1_\theta(X^s), \tag{1}$$
$$O_2, O'_2 = f^2_\theta(O_1, O'_1), \tag{2}$$
$$O, O' = f^3_\theta(O_2, O'_2). \tag{3}$$

Three loss functions guide the training of our feature extractor using synthetic data X^s. First, given the annotated labels of synthetic data, the network learns the classification loss function, which includes the cross-entropy function. This can be formulated as

$$L_{cls} = CE(f_\phi(O), y) + CE(f_\phi(O'), y), \tag{4}$$

where CE is the cross entropy function and y represent the annotated label corresponding to X^s.

Second, taking inspiration from [7], the authors propose the learnable consistency loss for test-time training domain generalization. The consistency loss has proven effective in reducing the domain gap between synthetic and real domains. We employ this learnable consistency loss as follows

$$L_{cont} = \| f_c(O - O') \|_2 . \tag{5}$$

f_c is the learnable parameter for aligning with classification loss. It can be represented as

$$\boldsymbol{g}_{cls} = \nabla_\theta(CE(f_\phi(O), y) + CE(f_\phi(O'), y)), \tag{6}$$
$$\boldsymbol{g}_{sub} = \nabla_\theta(\| f_c(O - O') \|_2), \tag{7}$$
$$L_c = \| \hat{\boldsymbol{g}}_{cls} - \hat{\boldsymbol{g}}_{sub} \|_2, \tag{8}$$

where $\hat{\boldsymbol{g}}(\cdot)$ denote the normalized gradients.

Third, to enhance feature representation, we introduce an auxiliary loss by solving temporal jigsaw puzzles. This approach [23] has demonstrated its effectiveness in VAD. It can formulated as

$$L_{aux} = \frac{1}{l_t}\sum_{l_t}^{1} CE(T_i, f_t(t_i)) \tag{9}$$

where T_i, t_i and f_t are the ground truth, predicted position outputs on X^s_{nor} input and temporal classifier, respectively. The alignment between the consistency and classification loss

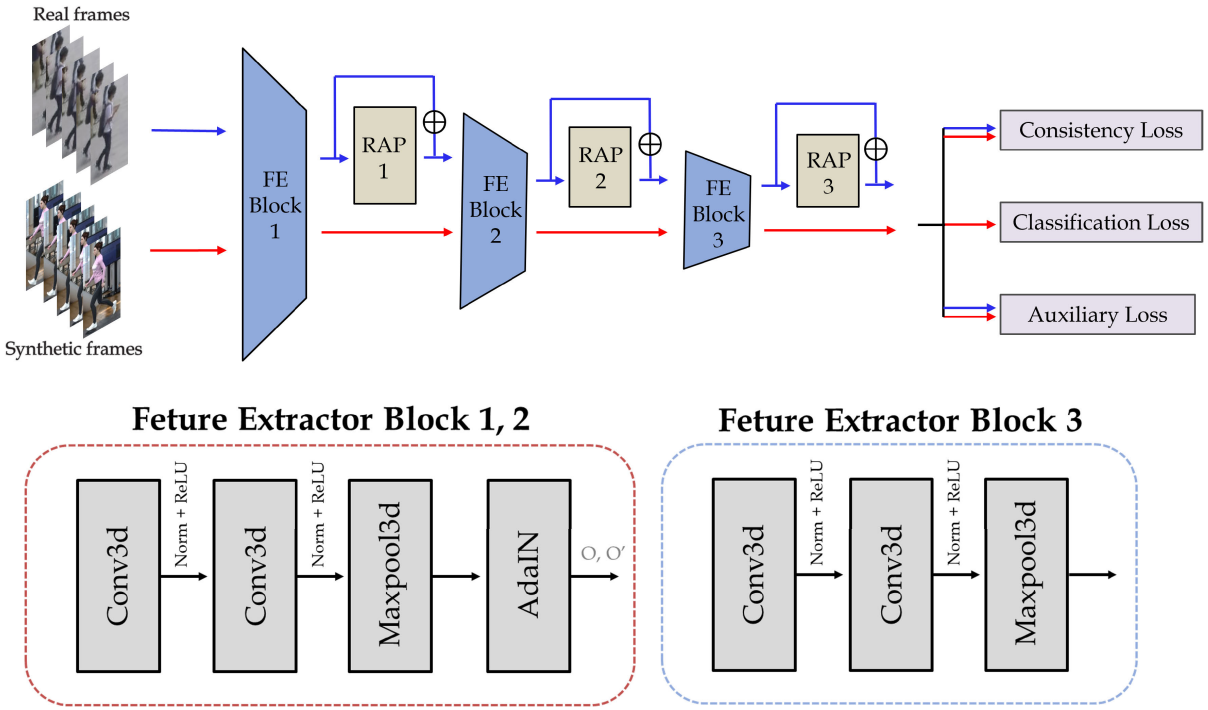

FIGURE 2. The architecture of SYRFA consists of a feature extractor and RAP block. Red arrow propagation is the stream of learning on synthetic knowledge (Phase 1) and Blue arrow propagation is the stream of adaptation on the real domain (Phase 2). In phase 1, the network is optimized on three losses. In phase 2, the network is trained on two losses. The feature extractor blocks 1, 2, and 3 are constructed almost the same but add the AdaIN technique on blocks 1, 2.

is crucial for guiding the feature representations during the adaptation to the real domain phase.

The networks of the feature extractor and classifier are optimized by minimizing the above losses, and this can be represented as

$$\min_{f_{\theta,\phi}} L_{cls} + L_{cont} + L_{aux}. \tag{10}$$

Additionally, the learnable parameters of consistency loss are trained by minimizing the L_c and can be formulated as

$$\min_{f_c} L_c. \tag{11}$$

The AdaIN augmentation technique is employed to extract features that emphasize content information. Utilizing these distinctive features, our feature extractor is trained to prioritize content features more robustly. This is achieved through the classification loss of O and O', which possess different scene information but similar content features. To impose more constrained learning, we introduce a consistency loss in conjunction with the classification loss. This additional loss function ensures close distances between O and O', guiding the learning process to align with the same direction as the classification loss. Furthermore, the feature extractor is configured to enhance feature representation by learning the temporal order of normal input. This additional configuration contributes to the overall robustness and temporal coherence of the feature extraction process.

B. ADAPTATION TO REAL DOMAIN

In this section, our primary objective is to align features between the real and synthetic domains. Optimizing on solely classification loss leads to easily overfitting problems due to accessing only normal real data. To mitigate the domain shift, we introduce a pretrained consistency loss, which is aligned with the classification loss on synthetic data and new additional parameters called RAP. Our RAP consists of the residual path unlike [7]. The residual architecture is used to learn local patterns [24]. In the VAD task, such local patterns are essential because circumstantial feature representation is acquired along with details of object motion. Therefore, we design RAP which is more suitable for VAD.

The weights of the pretrained feature extractor from the prior phase are kept fixed. The adaptation process is visualized as the blue arrow path in Fig. 2. The normal input of real data which is obtained from object-level frames by using an object detector is denoted as $X^r \in \mathbb{R}^{C\times T\times H\times W}$ where C, T, H, W represent channel size, temporal length, height and width, respectively. The RAP block and n-th RAP block are represented as f_τ and f_τ^n. The network for adaptation can formulated as

$$H_1, H_1' = M^1(X^r), \tag{12}$$
$$H_2, H_2' = M^2(H_1, H_1'), \tag{13}$$
$$H, H' = M^3(H_2, H_2'), \tag{14}$$
$$M^n(H_n) = f_\tau^n(f_\theta^n(H_n)). \tag{15}$$

Additionally, the weight of the RAP block has the same size as the input and is denoted as $w_\tau^n \in \mathbb{R}^{c\times t\times h\times w}$, where c, t, h, and w represent the channel size, temporal length, height, and width of the input to the RAP block. This can be formulated as

$$f_\tau^n(r) = r \times w_\tau^n + r, \tag{16}$$

where r and $\times$ are the input of RAP block and elementwise multiplication, respectively. The output of the RAP block, $f_\tau^n(r) \in \mathbb{R}^{c\times t\times h\times w}$, has the same size as the input because the RAP block activates the feature by elementwise multiplication and is designed to act as a residual path.

To avoid overfitting problems and promote rich feature representation, we apply two self-supervised loss functions during the adaptation phase on real data X^r. These loss functions are consistent with those from the prior phase: consistency and temporal auxiliary loss. The networks of the RAP block and the auxiliary classifier are optimized by minimizing the consistency and auxiliary losses, formulated as

$$\min_{f_{c,\tau}} L_{cont} + L_{aux}. \tag{17}$$

C. TOTAL LOSS FUNCTIONS AND INFERENCE

The proposed SYRFA framework is trained in two phases with a minimal number of additional parameters for adaptation, eliminating the need for large additional models such as VAE and Cycle-GAN. The total loss functions consist of the classification, consistency, and auxiliary losses. Our two-phase loss functions can be formulated as

$$L_{phase1} = \lambda_{cls}L_{cls} + \lambda_{cont}L_{cont} + \lambda_{aux}L_{aux}, \tag{18}$$

$$L_{phase2} = \lambda_{cont}L_{cont} + \lambda_{aux}L_{aux}. \tag{19}$$

Given unseen data, the input is first passed through the feature extractor, and then the output is processed through a classifier, resulting in a score

$$Score = f_\phi(f_\theta(X^{unseen})). \tag{20}$$

Following in [13], the instance-level anomaly scores are assembled into an anomaly map with the same shape as the input frame. The frame-level anomaly score is obtained by taking the maximum value in each frame of the anomaly map.

IV. EXPERIMENTS

A. EXPERIMENTAL SETUP

We evaluate the experimental results of our method on real datasets widely used in the VAD task.

UCSD Ped2 contains 16 training videos and 12 testing videos with fixed locations. It includes abnormal events in testing videos such as skateboarding, riding bikes and riding vehicles, etc. Each video has a resolution of 240 × 360 pixels in gray scale.

Shanghai Tech is a large-scale dataset that contains 330 training videos which contain only normal events and 107 testing videos which include both normal and abnormal events such as fighting, riding bikes, and robbery, etc. Additionally, It captured 130 abnormal scenarios with 13 different locations. The video frames are 480 × 856 resolutions.

1) EVALUATION METRIC

We employ the Area under ROC curve (AUC) with respect to the ground-truth annotations to evaluate the frame-level performance of our framework. Excellent anomaly detection method has a high AUC value. For AUC results, we first obtain the anomaly scores for all video frames and then calculate the scores globally for each dataset.

B. TRAINING DETAILS

To extract the object-level input frames, we employ the object detector named YOLOv3 which is pretrained on the MS COCO dataset. Following [6], the equivalent object detector, YOLOv3, is used which allows for fair comparison. We follow [13] to set the confidence thresholds which are 0.5 and 0.8 for Ped2 and Shanghai Tech during inference. For synthetic dataset supervised learning, we set the synthetic data confidence thresholds 0.5, and 0.9 for Ped2, Shanghai Tech. The cycle of phase 2 learning is executed per 20 and 10 training iterations for Ped2 and Shanghai Tech. The input size is $l_t \times 64 \times 64 \times 3$ where l_t is the length of frames. Adam optimizer is used with $\beta_1 = 0.9$ and $\beta = 0.999$. The learning rates for optimizing L_{phase1}, L_c and L_{phase2} are 1e-5, 1e-6 and 1e-6. The architecture of the feature extractor is the same as [13] and we finetune the pretrained network [23] on each VAD dataset.

C. ANOMALY DETECTION RESULTS

In Table 1, we provide a comprehensive comparison of our SYRFA framework with state-of-the-art (SOTA) methods [5], [6], [12], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [37], [38], [39] in terms of frame-level AUC (%). We further categorize the SOTA methods into four categories, including reconstruction and prediction-based methods, which have dominated the field of VAD. Among the SOTA methods, [6] and [36] leverage the object-level input frames through FPN and YOLOv3 object detectors, respectively. The others employ the resized frame-level input frames without object detectors.

1) RESULT ON UCSD PED2

UCSD Ped2 is one of the most popular benchmarks of VAD. Most SOTA methods achieve over 90% AUC on this dataset, and the differences between these methods are relatively small. Our SYRFA is slightly lower than the method but shows correspondingly good results when compared to SOTA methods.

2) RESULT ON SHANGHAITECH

The ShanghaiTech dataset is a large-scale dataset encompassing various scenes and events. As presented in Table 1, our SYRFA surpasses SOTA methods with an AUC of 84.6%, outperforming previous methods by a margin of 0.8%. Unlike

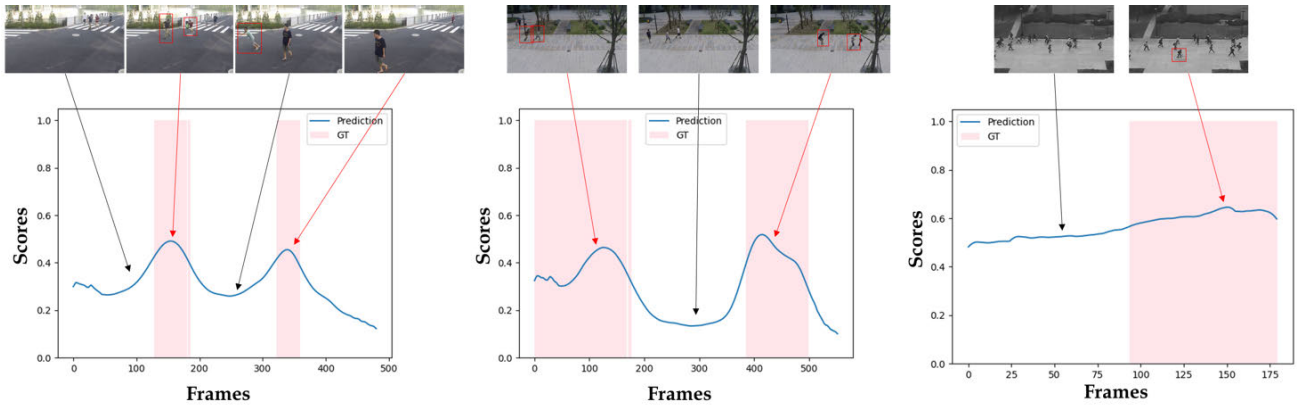

FIGURE 3. Frame-level scores and anomaly localization examples for a test video from the Shanghai Tech and Ped2 dataset. The blue line is anomaly scores and the red bar plot means the corresponding ground truth. The example is 07_0006 test video from the Shanghai Tech dataset (Left). The example is the 03_0061 test video from the Shanghai Tech dataset (Middle). The example is 02 test video from the Ped2 dataset (Right).

UCSD Ped2, Shanghai Tech contains multiple scenes and diverse anomaly types [5]. In particular, our SYRFA performs significantly better than other SOTA methods on the Shanghai Tech, indicating that the proposed method generalizes well when applied to a real-world environment.

D. ABLATION STUDIES

1) COMPONENTS ANALYSIS

To evaluate the contributions of different components in our proposed SYRFA framework, we conducted a series of ablation experiments on the ShanghaiTech dataset. In Table 2, we present the results of these experiments.

- **Experiment 1:** This experiment uses only the classification loss from Phase 1 and achieves an AUC of 83%.
- **Experiment 2, 3:** By adding separately both temporal auxiliary loss and consistency loss from Phase 1, the performance increases by 0.2% and 0.3%, respectively.
- **Experiment 4:** This experiment incorporates both temporal auxiliary loss and consistency loss, resulting in significant performance improvement, highlighting the importance of combining these two losses.
- **Experiment 5:** In the comparative analysis between Adaptive Parameters (AP) [7] and our proposed RAP, we conduct experiments involving AP. The obtained results reveal an increase, yet the performance remains suboptimal.
- **Experiment 6:** In this experiment, we introduce the Residual Additional Parameters (RAP) component, which plays a crucial role in reducing the domain gap. When all parts are combined, SYRFA achieves the best performance with an AUC of 84.6%. Particularly, the observed improvement in results demonstrates the effectiveness of the proposed framework and RAP.

2) HYPERPARAMETER ANALYSIS

The overall loss function contains three hyperparameters, called λ_{cls}, λ_{cont} and λ_{aux}. We explore the optimal hyperparameter setting and report the results in Table 4. Experiment 1 represents the outcome of a standard experiment with all hyperparameters set to 1. Experiments 2-3 are conducted by modifying only hyperparameter λ_{cont}, while Experiments 4-6 are executed by altering only the hyperparameter λ_{aux}. Subsequently, Experiments 7-9 involve various combinations of the modified λ_{cont} and λ_{aux}. From the experimental results, it becomes evident that while the outcomes distinctly enhance performance depending on λ_{aux}, adjusting λ_{cont} produces poor results. Consequently, the hyperparameters corresponding to Experiment 6, where only the λ_{aux} is modified, are implemented. This yields a result that is 1.5% higher than the outcome of the standard Experiment 1.

3) AUPRC AND MAX-F1 ANALYSIS

To elaborate comparison with SOTA methods [12], [31], We investigate additional evaluation metrics such as Area Under the Precision-Recall Curve (AUPRC) score and Max-F1 score which is the harmonic mean of precision and recall. For the Max-F1 score, we report the maximum F1 score from the results of all thresholds. Likewise AUC metric, We aggregate the anomaly scores of all video frames. A higher score (AUPRC and F1) indicates better anomaly detection performance. As shown in Tab. 3, the performance of our SYRFA on AUPRC and Max-F1 scores else except Max-F1 in ShanghaiTech achieve significant results. This demonstrates that our method is a more effective framework from AUC as well as AUPRC and F1 metric perspectives. For the ShaghaiTech, the performance of our SYRFA does not show optimal results. Because the ShanghaiTech includes various scenes, the combinations between object contents and backgrounds are complicated. However, our SYRFA achieves competitive compared with other methods.

4) COMPARISON WITH EXISTING FEATURE ALIGNMENT METHOD

Our SYRFA framework is designed to alleviate domain shift through feature alignment. Consequently, it is crucial

TABLE 1. Comparison with state-of-the-art methods in terms of micro-AUROC (%). The best and second-best results are shown in bold and underlined, respectively. The methods are divided into four categories which contain reconstruction, prediction, synthetic, and others.

Category	Method	Ped2	STC
Reconstruction	Luo et al. [25]	92.2	68.0
	Gong et al. [26]	94.1	71.2
	Park et al. (Recon.) [27]	90.2	69.8
	Chang et al. [28]	96.5	73.3
	Astrid et al. (Patch based) [12]	94.8	72.5
	Astrid et al. (Skip frame based) [12]	96.5	76.0
	Wang et al. [29]	**97.7**	71.3
Prediction	Liu et al. [14]	95.4	72.8
	Lee et al. [30]	96.6	76.2
	Wang et al. [31]	96.3	76.6
	Dong et al. [32]	95.6	73.7
	Park et al. (Pred.) [27]	97.0	70.5
	Huang et al. [33]	95.5	76.5
Others	Ye et al. [34]	96.8	73.6
	Tang et al. [35]	96.3	73.0
	Ionescu et al. [36]	94.3	78.7
	Wu et al. [37]	96.9	–
	Liu et al. [38]	<u>97.6</u>	77.6
	Shi et al. [39]	<u>97.6</u>	78.8
Synthetic	Acsintoae et al. [5]	–	83.7
	Liu et al. [6]	–	<u>83.8</u>
	SYRFA (ours)	96.9	**84.6**

to compare our feature alignment methodology with other existing methods to substantiate its effectiveness. Liu et al. [6] primarily focus on generating abnormal samples for VAD, with feature alignment applied incidentally through the use of the Gradient Reversal Layer (GRL). In our SYRFA, we incorporate GRL with domain labels to substitute the

TABLE 2. The results on ShanghaiTech for components experiments.

EXP	Phase1			Phase2				AUC
	Cls	Aux	Cont	Aux	Cont	AP	RAP	
1	✓							83.0%
2	✓	✓		✓				83.2%
3	✓		✓		✓			83.3%
4	✓	✓	✓	✓	✓			84.3%
5	✓	✓	✓	✓	✓	✓		84.5%
6	✓	✓	✓	✓	✓		✓	84.6%

TABLE 3. AUPRC and Max-F1 scores of the anomaly detection results.

	Ped2		ShanghaiTech	
Score	AUPRC	Max-F1	AUPRC	Max-F1
Wang et al. [31]	–	–	72.4%	74.5%
Astrid et al. [12]	79.6%	79.1%	76.6%	**77.3%**
SYRFA(ours)	**99.4%**	**96.4%**	**80.9%**	75.1%

remaining loss while preserving the classification loss. The experimental results, presented in Table 5, demonstrate the significant performance of our SYRFA across all datasets. This finding underscores the limitations of relying solely on feature alignment through GRL, emphasizing the necessity for a more detailed feature alignment approach. Therefore, SYRFA has the ability to learn domain-invariant representations through advanced feature alignment without generating abnormal samples. Also, This implies a reduction in domain shift when applying synthetic knowledge to the real data domain.

5) FEATURE DISTANCE ANALYSIS

We conducted an experiment measuring the L2 distance between the original input and augmented input using AdaIN, which is a part of our proposed RAP. This experiment aims to demonstrate the effectiveness of RAP in handling the domain gap. As shown in Fig. 4, RAP results in a smaller distance compared to previous adaptive parameter approaches on VAD datasets. This finding suggests that RAP significantly contributes to obtaining transferable feature representations and effectively addressing domain gaps.

E. VISUALIZATION

To validate our approach, we generated anomaly scores for the test data and conducted visualizations of frame-level anomaly scores as well as an example of anomaly localization. We represent the results of this validation using ShanghaiTech and Ped2 datasets in Fig. 3.

In the three examples provided, it is evident that the output scores for abnormal events consistently exhibit higher values than the output scores for normal events. This distinction in output scores highlights the discriminative power of our SYRFA framework in effectively identifying anomalies within video data.

These visualizations provide a qualitative assessment of the effectiveness of our approach in detecting anomalies within video data, demonstrating its potential for real-world applications in video anomaly detection.

TABLE 4. The results on ShanghaiTech of hyperparameter experiments.

EXP	λ_{cls}	λ_{cont}	λ_{aux}	AUC
1	1	1	1	83.1%
2	1	0.1	1	83.2%
3	1	0.01	1	83.5%
4	1	1	0.1	84.4%
5	1	1	0.01	84.5%
6	1	1	0.05	**84.6%**
7	1	0.01	0.05	83.3%
8	1	0.1	0.05	83.4%
9	1	0.5	0.05	84.1%

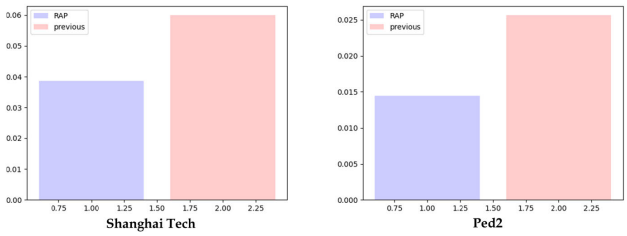

FIGURE 4. A comparison of additional parameters methods between our proposed RAP and [7]. The blue bar is the L2 distance of features with the proposed RAP and the red bar means the L2 distance of features with [7].

TABLE 5. The experiments for comparison feature alignment methods.

Methods	STC	Ped2
GRL	81.4	96.1
SYRFA (ours)	**84.6**	**96.9**

V. LIMITATIONS

While our framework has shown improvements without relying on a generation model, we acknowledge that there are certain limitations in addressing the complexities of the VAD task. VAD exhibits a scene-dependent nature, implying that what may be considered anomalous in one scene, such as riding a bicycle on a pedestrian road, could be entirely normal in another, like a dedicated bicycle lane. However, our current feature extractor has been primarily designed to emphasize the acquisition of content-related features, primarily through the use of the AdaIN technique. These results can be

inferred from the Max-F1 score for the ShanghaiTech dataset. Furthermore, the application in real-world environments, characterized by a wider range of diverse scenes, may lead to misjudgments or errors in anomaly detection. Consequently, our future work will be dedicated to extract scene-specific information by eliminating content information from input features within our framework.

VI. CONCLUSION

In this paper, our objective is to attain a domain-invariant representation, facilitating the application of knowledge derived from synthetic data to real-world scenarios. We address this challenge through cross-domain feature alignment. Our proposed framework comprises two key stages: ''Learning on Synthetic Knowledge'' and ''Adaptation to Real Domain.'' In the former, we leverage rich data and annotated labels from synthetic sources. The latter involves the network learning domain-invariant representations in a self-supervised manner, aimed at mitigating domain shift when exposed to real data. Additionally, we introduced Residual Additional Parameters (RAP), specifically designed for enhanced transferability in the Video Anomaly Detection (VAD) task. As a result, our feature alignment-based method, which does not rely on additional generation models, has shown effectiveness in the VAD dataset. However, our proposed framework exhibited limitations as it excluded scene-specific information, focusing solely on content-related information. In our future work, we plan to explore methodologies that obtain scene-related feature representations by eliminating content information from input features and introducing both content and scene information.

REFERENCES

[1] S. Park, C. J. Cho, B. Ku, S. Lee, and H. Ko, ''Compact HF surface wave radar data generating simulator for ship detection and tracking,'' *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 6, pp. 969–973, Jun. 2017.

[2] J. Seo and H. Ko, ''Face detection using support vector domain description in color images,'' in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, May 2004, p. 729.

[3] S. C. Byun, D. B. Choi, B. H. Ahn, and H. Ko, ''Traffic incident detection using evidential reasoning based data fusion,'' in *Proc. World Congr. Intell. Transport Syst. (ITS)*, Toronto, ON, Canada, Nov. 1999.

[4] W. Zhang, H. Sun, D. Zhao, L. Xu, X. Liu, H. Ning, J. Zhou, Y. Guo, and S. Yang, ''A streaming cloud platform for real-time video processing on embedded devices,'' *IEEE Trans. Cloud Comput.*, vol. 9, no. 3, pp. 868–880, Jul. 2021.

[5] A. Acsintoae, A. Florescu, M.-I. Georgescu, T. Mare, P. Sumedrea, R. T. Ionescu, F. S. Khan, and M. Shah, ''UBnormal: New benchmark for supervised open-set video anomaly detection,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20111–20121.

[6] Z. Liu, X.-M. Wu, D. Zheng, K.-Y. Lin, and W.-S. Zheng, ''Generating anomalies for video anomaly detection with prompt-based feature mapping,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 24500–24510.

[7] L. Chen, Y. Zhang, Y. Song, Y. Shan, and L. Liu, ''Improved test-time adaptation for domain generalization,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 24172–24182.

[8] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, ''Robust real-time unusual event detection using multiple fixed-location monitors,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 555–560, Mar. 2008.

[9] J. Kim and K. Grauman, ''Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2921–2928.

[10] Y. Cong, J. Yuan, and J. Liu, ''Sparse reconstruction cost for abnormal event detection,'' in *Proc. CVPR*, Jun. 2011, pp. 3449–3456.

[11] K. Kim and H. Ko, ''Hierarchical approach for abnormal acoustic event classification in an elevator,'' in *Proc. 8th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2011, pp. 89–94.

[12] M. Astrid, M. Z. Zaheer, J.-Y. Lee, and S.-I. Lee, ''Learning not to reconstruct anomalies,'' 2021, *arXiv:2110.09742*.

[13] M.-I. Georgescu, A. Barbalau, R. T. Ionescu, F. S. Khan, M. Popescu, and M. Shah, ''Anomaly detection in video via self-supervised and multi-task learning,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12737–12747.

[14] W. Liu, W. Luo, D. Lian, and S. Gao, ''Future frame prediction for anomaly detection—A new baseline,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6536–6545.

[15] Y. Fan, G. Wen, D. Li, S. Qiu, M. D. Levine, and F. Xiao, ''Video anomaly detection and localization via Gaussian mixture fully convolutional variational autoencoder,'' *Comput. Vis. Image Understand.*, vol. 195, Jun. 2020, Art. no. 102920.

[16] G. Yu, S. Wang, Z. Cai, E. Zhu, C. Xu, J. Yin, and M. Kloft, ''Cloze test helps: Effective video anomaly detection via learning to complete video events,'' in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 583–591.

[17] B. Lee and H. Ko, ''Injecting sparsity in anomaly detection for efficient inference,'' in *Proc. 17th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Nov. 2021, pp. 1–5.

[18] Y. Jin, J. Hong, D. Han, and H. Ko, ''CPNet: Cross-parallel network for efficient anomaly detection,'' in *Proc. 17th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Nov. 2021, pp. 1–8.

[19] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, ''Learning temporal regularity in video sequences,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 733–742.

[20] W. Luo, W. Liu, and S. Gao, ''Remembering history with convolutional LSTM for anomaly detection,'' in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 439–444.

[21] X. Feng, D. Song, Y. Chen, Z. Chen, J. Ni, and H. Chen, ''Convolutional transformer based dual discriminator generative adversarial networks for video anomaly detection,'' in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 5546–5554.

[22] X. Huang and S. Belongie, ''Arbitrary style transfer in real-time with adaptive instance normalization,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1510–1519.

[23] G. Wang, Y. Wang, J. Qin, D. Zhang, X. Bao, and D. Huang, ''Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles,'' in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Oct. 2022, pp. 494–511.

[24] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, ''Residual dense network for image super-resolution,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.

[25] W. Luo, W. Liu, and S. Gao, ''A revisit of sparse coding based anomaly detection in stacked RNN framework,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 341–349.

[26] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. Van Den Hengel, ''Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1705–1714.

[27] H. Park, J. Noh, and B. Ham, ''Learning memory-guided normality for anomaly detection,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14360–14369.

[28] Y. Chang, Z. Tu, W. Xie, and J. Yuan, ''Clustering driven deep autoencoder for video anomaly detection,'' in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 329–345.

[29] L. Wang, J. Tian, S. Zhou, H. Shi, and G. Hua, ''Memory-augmented appearance-motion network for video anomaly detection,'' *Pattern Recognit.*, vol. 138, Jun. 2023, Art. no. 109335.

[30] S. Lee, H. G. Kim, and Y. M. Ro, ''BMAN: Bidirectional multi-scale aggregation networks for abnormal event detection,'' *IEEE Trans. Image Process.*, vol. 29, pp. 2395–2408, 2020.

[31] X. Wang, Z. Che, B. Jiang, N. Xiao, K. Yang, J. Tang, J. Ye, J. Wang, and Q. Qi, ''Robust unsupervised video anomaly detection by multipath frame prediction,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 6, pp. 2301–2312, Jun. 2022.

[32] F. Dong, Y. Zhang, and X. Nie, "Dual discriminator generative adversarial network for video anomaly detection," *IEEE Access*, vol. 8, pp. 88170–88176, 2020.
[33] X. Huang, C. Zhao, C. Gao, L. Chen, and Z. Wu, "Synthetic pseudo anomalies for unsupervised video anomaly detection: A simple yet efficient framework based on masked autoencoder," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.
[34] M. Ye, X. Peng, W. Gan, W. Wu, and Y. Qiao, "AnoPCN: Video anomaly detection via deep predictive coding network," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1805–1813.
[35] Y. Tang, L. Zhao, S. Zhang, C. Gong, G. Li, and J. Yang, "Integrating prediction and reconstruction for anomaly detection," *Pattern Recognit. Lett.*, vol. 129, pp. 123–130, Jan. 2020.
[36] R. T. Ionescu, F. S. Khan, M.-I. Georgescu, and L. Shao, "Object-centric auto-encoders and dummy anomalies for abnormal event detection in video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7834–7843.
[37] P. Wu, J. Liu, and F. Shen, "A deep one-class neural network for anomalous event detection in complex scenes," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 7, pp. 2609–2622, Jul. 2020.
[38] Y. Liu, Z. Xia, M. Zhao, D. Wei, Y. Wang, S. Liu, B. Ju, G. Fang, J. Liu, and L. Song, "Learning causality-inspired representation consistency for video anomaly detection," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 203–212.
[39] C. Shi, C. Sun, Y. Wu, and Y. Jia, "Video anomaly detection via sequentially learning multiple pretext tasks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 10330–10340.

JONGHWAN HONG (Student Member, IEEE) received the B.S. degree in mechanical engineering from Sejong University, Seoul, South Korea, in 2020. Since 2020, he has been the integrated M.S. and Ph.D. program in electrical engineering with Korea University, Seoul. His research interests include video anomaly detection, video analysis, and deep neural networks.

BOKYEUNG LEE received the B.S. degree in electronic engineering from Kwangwoon University, Seoul, South Korea, in 2018. Since 2019, he has been the integrated M.S. and Ph.D. program in electrical engineering with Korea University, Seoul. His research interests include domain generalization and knowledge distillation on image and video processing.

KYUNGDEUK KO (Graduate Student Member, IEEE) received the B.S. degree in biomedical engineering from Yonsei University, in 2017, and the M.S. degree in electrical engineering from Korea University, in 2019, where he has been pursuing the Ph.D. degree, since 2020. His research interests include computer vision, biomedical signal/image processing, and artificial intelligence.

HANSEOK KO (Senior Member, IEEE) received the B.S. degree in electrical engineering from Carnegie Mellon University, in 1982, the M.S. degree in electrical engineering from Johns Hopkins University, in 1988, and the Ph.D. degree in electrical engineering from CUA, in 1992. In March 1995, he joined the Faculty of the Department of Electronics and Computer Engineering, Korea University, where he is currently a Professor. At the onset of his career, he was with WOL, MD, USA, where his work involved signal and image processing.