Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 14, No. 3, 2025

From Pixel Fidelity to Task Performance: Image Quality Challenges and Paradigm Shift in Public Security Video Surveillance

Guochen Yang *, Zicheng Wang

Department of information network security, People's Public Security University of China, Beijing, China
*** Corresponding author:** Guochen Yang (Email: 19895608786@163.com)

Abstract: As public safety video surveillance enters the era of intelligence, image quality has become a bottleneck limiting the performance of visual perception systems. This paper reviews the development of image quality research in this field: from the early focus solely on pixel-level fidelity evaluation to evolving into evaluation guided by high-level visual task performance. In this paper, we first analyze the image quality challenges faced in public security scenarios. Then, focusing on how to ensure the perception reliability of intelligent systems, we summarize three technical approaches: first, using image preprocessing technologies such as image enhancement, restoration, and super-resolution to directly improve input quality; second, designing targeted detection and recognition models to enhance system robustness against quality degradation; third, leveraging multimodal information fusion, such as visible-infrared integration, to compensate for the limitations of a single modality. This review aims to clarify that the core of image quality research has shifted toward building a task-oriented and degradation-insensitive robust visual perception system.

Keywords: Image Quality Evaluation; Public Security Surveillance; Image Enhancement; Multi-modal Fusion.

1. Introduction

1.1. Intelligent Transformation of Public Security Video Surveillance

The public security video surveillance system is undergoing a rapid transformation from traditional recording to intelligent stage. The breakthrough of deep learning in computer vision has enabled large-scale deployment of functions such as target detection, face recognition, person re-identification, and behavior understanding in urban surveillance networks, thus becoming an important part of smart city governance infrastructure [1-3].

In this process, the image captured by the camera constitutes the basic input of the intelligent sensing link, and its quality directly affects the accuracy, recall rate and stability of the back-end algorithm. Existing research has clearly pointed out that the decline in the quality of the input image will lead to non-linear performance degradation in multiple visual tasks. For example, in the target detection task, poor image quality may lead to missed detection of targets ; in the person re-identification (Re-ID) task, matching errors may occur ; in face recognition, it may lead to a significant reduction in the stability of facial feature extraction [4]. Therefore, image quality is a key factor restricting the reliability of intelligent monitoring system.

1.2. Imaging Complexity in Real Surveillance Environment

Different from natural images or laboratory data, the changing environment and scene of public safety monitoring lead to a variety of degradation of images. Day and night illumination changes, low illumination, backlight conditions, scattering degradation caused by rain and fog weather, small target problems caused by long distance, motion blur caused by fast movement, and parameter differences between cameras may cause serious degradation of image quality [5, 6]. These degradation phenomena often work together in the form of superposition, making the ideal image difficult to obtain ; at the same time, they are highly variable in space and time, making it difficult for the back-end model to establish a stable representation. In addition, the rise of deep forgery technology in recent years has further brought about risks at the semantic level, making ' image authenticity ' an extended dimension of image quality [7]. In summary, the complexity of public safety monitoring scenarios requires a systematic understanding of the imaging degradation mechanism, and an intelligent vision system that can adapt to ' non-ideal input ' is urgently needed.

1.3. Evolution of Image Quality Concepts

Traditional image quality assessment mainly relies on PSNR, SSIM and other reference indicators to measure the pixel deviation between the image and the ideal image. However, such indicators face many fundamental problems in the monitoring environment. Firstly, public safety surveillance images usually do not have ideal non-degraded reference images ; secondly, there is no stable correlation between pixel-level deviation and semantic-level task performance [8] ; a large number of studies have shown that some ' higher quality ' images on visual experience do not improve detection and recognition performance, while some images with lower objective indicators can maintain stable performance at the task level. [9]. Therefore, the understanding of image quality in the monitoring field has gradually shifted from ' visual fidelity ' to ' task availability ', and a new standard has been introduced : image quality must serve high-level tasks such as target detection, face recognition and ReID, rather than stay at the pixel level optimization. [10].

In this context, the research focus of image quality improvement has also changed, gradually evolving from

simple pixel enhancement to model robustness and whole process collaborative optimization. Specifically, the research directions mainly include : first, image enhancement and restoration technology, which aims to effectively compensate and repair degraded input images ; the second is to optimize the model design to improve the robustness of the model to degradation conditions, so as to learn and extract stable feature representations directly under non-ideal image conditions. The third is multi-modal fusion technology, such as combining infrared and visible images to use complementary information of different modalities to break through the limitations of a single modality in harsh environments (such as low light or bad weather) [11-13]. The research in these directions has jointly promoted the systematic evolution of image quality research system in the field of public safety from traditional ' pixel-driven ' to more advanced ' task-driven '.

1.4. Research Motivation and Paper Structure

With the large-scale deployment of intelligent monitoring systems in real urban environments, building a visual model that can maintain stable performance under complex conditions has become an urgent requirement in the field of public security combat. Based on this demand, this paper reviews the development status of image quality evaluation criteria and model optimization from the perspective of the influence of image quality on intelligent perception tasks.

The second chapter will analyze the main imaging challenges in the surveillance scene as the basis for understanding the subsequent technical paths. The third chapter systematically discusses the image quality assessment methods in the field of public safety. The fourth chapter introduces the research of deep learning model in dealing with degraded images and improving the robustness of the model. The fifth chapter summarizes the progress of multimodal fusion technology in breaking through the limitation of visible light imaging. The sixth chapter discusses the task-oriented quality evaluation system and system-level optimization direction based on the research trend.

2. Imaging Challenges and Quality Bottlenecks in Public Safety Scenarios

The environment captured by the public safety monitoring system is highly complex and uncontrollable. Due to the regular changes of sunshine and the diversity of monitoring scenes, the imaging conditions change frequently. A variety of degradation causes are superimposed on the monitoring image. Compared with the research dataset, video surveillance images are often difficult to obtain ideal images, extract stable features, and maintain consistency in semantic analysis due to lighting conditions, meteorological differences, scene changes, and dynamic blur in the scene. This chapter will discuss the difficulties faced by public security video surveillance cameras from the aspects of lighting conditions, weather conditions, detection target limitations, and cross-device differences.

2.1. Illumination Change and Influence of Low Illumination Conditions

In the scene of video surveillance in the field of public security, the lighting conditions change with time, weather and space environment. Under the condition of night or dark area, the surveillance camera works under extremely low illumination, resulting in a significant decrease in the signal-to-noise ratio of the image, and the detail texture is submerged by noise. The dynamic range of the camera is limited, making it difficult to extract key features such as target edge, color, and material. [14]. In extreme dark scenes, the image sensor needs to extend the exposure time to make enough light pass through the lens, which is easy to cause motion blur or noise accumulation. [15] In addition, the surveillance camera often operates with a fixed exposure strategy, and does not adaptively adjust the aperture, ISO, and shutter speed according to the scene like a consumer-grade camera, making the video surveillance camera less adaptable to night scenes.

Another scene that the video surveillance camera is difficult to deal with is the backlight scene. In areas such as indoor and outdoor junctions, tunnel entrances, and building entrances, the light ratio is often extremely high. When the pedestrian 's face falls into shadow, due to the limited dynamic range of the camera sensor, the details of the extremely bright and extremely dark areas cannot be clearly captured at the same time. Under backlight, in order not to be exposed, the camera will automatically reduce the overall exposure, resulting in dark objects becoming dark and losing details due to insufficient exposure [16].

2.2. The Influence Mechanism of Severe Weather and Environmental Degradation

Severe weather is one of the main degradation factors of public safety scenarios. Rain, snow, fog, haze and other weather factors scatter and absorb light, resulting in a decrease in the overall contrast of the image, blurred distant targets, and blurred details. [17] Under this condition, pedestrians or vehicles are almost completely blurred into low-frequency objects at medium and long distances, and the detection algorithm is difficult to distinguish their boundaries.

Rain and snow weather will also introduce dynamic interference : raindrops or snowflakes form bright scattering spots in the imaging process, block local areas, and form irregular reflections. These noises will seriously affect the stability of convolutional networks or Transformer models [23]. Compared with static haze, the dynamic occlusion caused by rain and snow has time-varying characteristics, which makes it difficult for the model to learn consistent representation in a short window. In addition, the high humidity environment may also reduce the transparency of the lens and reduce the luminous flux of the lens, which in turn affects the image quality.

2.3. Challenges Posed by Target Scale, Motion, and Occlusion

The scale of the target itself, the dynamic behavior and the multi-object spatial relationship of the scene also lead to imaging distortion. Since the camera is mostly installed at a high position and is far from the main monitoring area, the proportion of pixels in the image of the target is often very limited. For key targets such as pedestrians and vehicles, insufficient resolution directly leads to the lack of identifiable features. For example, in the person re-identification task, key cues such as clothing texture, backpack shape, and gait posture are difficult to accurately capture under remote imaging [18]. Studies have shown that when the number of target pixels is less than a certain threshold, the detection and recognition confidence of the visual model will decrease exponentially, and even a stable representation cannot be

established.

Dynamic ambiguity caused by motion is another kind of common degradation. Under low illumination conditions, the camera tends to extend the exposure time, which causes the fast-moving target to be smeared or faded in the image, thereby damaging the shape and contour information. [15]

The occlusion phenomenon in the video (such as crowd intersection, mutual occlusion of vehicles or partial occlusion of pedestrians by trees / cylinders) also leads to the lack of local semantic information. It is difficult for the detection and recognition model to recover the identity representation from incomplete data, resulting in recognition failure. [19].

2.4. Representation Inconsistencies Due to Cross-Device and Cross-Scene Variations

Due to differences in hardware, installation location, viewing angle, and internal imaging pipelines between cameras in the surveillance network, there are significant differences in the ' imaging style ' of cross-device images. [20] For example, the same target may exhibit completely different appearance under different cameras. This difference is particularly prominent in person re-identification and cross-camera tracking, which is one of the core factors leading to unstable characterization.

In addition, scene background differences can also cause significant domain shifts. Such as outdoor streets, underground garages, shopping mall halls, corridor entrances and other background complexity, lighting conditions and perspective changes are very different, making it difficult for visual models to maintain consistent representation in a unified feature space. [21]. Deep learning has strong dependence on training data and limited cross-scene generalization ability, which makes the public safety system often face the problem of ' the model is effective in the original scene and the performance is degraded in the new scene ' in the actual landing.

3. Image Quality Evaluation Methods for Public Security Video Surveillance

In the public security monitoring system, image quality not only directly affects the performance of visual perception algorithms, but also relates to the reliability of event detection, abnormal behavior recognition and intelligent analysis. Different from general photography or multimedia scenes, image quality assessment in surveillance scenes needs to take into account the degradation of the real environment, dynamic changes, and downstream task effectiveness. This chapter systematically discusses the image quality of public security video surveillance from four aspects : evaluation index system, subjective and objective evaluation methods, task-related evaluation and cross-modal quality evaluation.

3.1. Image Quality Assessment Metrics

Image quality assessment (IQA) is usually divided into three categories : subjective evaluation, objective index and task-driven evaluation [24]. Subjective evaluation obtains visual quality through manual observation, such as MOS (Mean Opinion Score) and DMOS (Difference MOS). Studies have shown that subjective evaluation can most directly reflect the perception preferences of human observers, but it is not suitable for large-scale real-time monitoring systems because of its high cost and difficulty in large-scale implementation [25].

Objective indicators predict image quality through mathematical models or signal processing methods, which can be divided into three categories : full reference (FR), no reference (NR) and half reference (RR). For example, PSNR, SSIM and MS-SSIM belong to the full reference index, and the pixel value operation of the distorted image and the non-degraded original image is used to calculate the distortion degree of the image. It has high reliability when there is a clear reference image, but the video surveillance image does not have the original image that can be considered as non-degraded, so it is not suitable for this method [26] ; bRISQUE, NIQE, PIQE and other non-reference methods are more suitable for monitoring scenarios [27].These indicators estimate the quality through statistical features or natural image priors, and rank the quality of a group of images, which has high practicability in video surveillance image quality evaluation.

Traditional signal-based evaluation methods are difficult to reflect the real impact of image quality on task performance, so deep learning models are gradually widely used in public safety systems. The task-related quality assessment method describes the image quality as the performance changes of tasks such as target detection, person re-identification, and license plate recognition. This method is more intuitive and more in line with the actual combat scene of public security [31]. By analyzing the correlation between specific task performance and image degradation types, the task-related IQA method skips the intermediate variable of pixel fidelity, and directly uses task-related evaluation accuracy and recall rate to represent the actual impact of different degradation types on the monitoring system, and to guide the hardware deployment and algorithm optimization direction [32].

3.2. Cross-modal Image Quality Assessment

Modern public safety systems increasingly integrate multi-modal sensors such as visible light, infrared, and millimeter waves to improve all-weather and multi-scene recognition capabilities. This is a cross-modal fusion image evaluation thinking. Its core is to use different modes and intervals to construct a unified evaluation framework that can monitor multiple scenes through appropriate weight distribution.

Taking cameras with different wavelengths of receiving light as examples, infrared images can provide stable contour information in low-light environment, but the texture details are insufficient. Visible light images provide rich texture and color information in well-illuminated scenes, but the signal-to-noise ratio is significantly reduced in low-light environments. Zhang et al. [34] proposed a method based on joint feature modeling and multi-modal scoring mechanism, which can evaluate different modal images uniformly, and can obtain the modal selection weights of different conditions through training.

At present, cross-modal evaluation still has systematic differences such as sensor calibration error and spatial mapping deviation of various monitoring scenarios, which is an important research direction in the current monitoring system design.

3.3. Summary

This chapter systematically combs the image quality evaluation methods in public security video surveillance. Firstly, it introduces the subjective and objective evaluation of traditional signal level and its application in public security

video surveillance, and then introduces the task-driven and cross-modal evaluation system. It is concluded that no-reference quality indicators are more suitable for complex environmental monitoring scenarios ; the image evaluation based on the recognition accuracy and recall rate of the final task can directly reflect the actual needs of the intelligent system in the field of public security, and the cross-modal quality evaluation.

4. Robustness Enhancement Strategy of Visual Model

Relying solely on image preprocessing, it is difficult to achieve multi-source degradation of the real world, and may introduce new artifacts or cause computational difficulties. Therefore, researchers have gradually shifted their focus from image preprocessing technology to enhancing the robustness of the model. By optimizing the model itself, it can stably obtain task-related features from the complex degraded images of public security scenes, so that it can automatically identify and filter common distortions such as significant noise, blur, low illumination, and occlusion. The core idea of this path is to make the model no longer rely on high-quality input through the optimization of network structure, feature learning strategy and training method, so that it can adapt to the general quality image of video surveillance camera and complete the task of person re-identification.

4.1. Robust Feature Learning in Detection Models

In the target detection task, scale change, small target distribution, local blur and complex background have a significant impact on the stability of the model. Lin et al. [35] used the unique pyramid network (FPN) of convolutional neural network (CNN) to integrate the deepest semantically rich information with higher-level high-resolution information with the top-down path and horizontal connection of countries, making it suitable for small target detection problems common in long-distance monitoring scenarios. On the COCO dataset, FPN has increased the accuracy of small target detection by 8 percentage points.

Single-stage detectors (such as YOLO series) have advantages in structural lightweight and inference speed. This is because they regard target detection as a single regression problem, directly predict bounding boxes and category probabilities on the feature map, omit the resampling step, achieve a balance between speed and accuracy, and adapt to the needs of actual scenes. It also proposes a multi-scale prediction method : detection on three feature maps with different resolutions can adapt to different scales of targets. This multi-scale prediction idea becomes the standard for subsequent detectors. [36].

Based on Transformer 's detection model, Carion et al. [37] proposed such as DETR. For the first time, the Transformer framework is applied to the target detection task, and the self-attention mechanism is used to model the long-range dependence, which shows stronger robustness when dealing with complex backgrounds, weak texture regions and partially occluded targets.

These structural improvements show that improving the robustness of the detection model does not depend on the preprocessing of the input image, and it is also a feasible path to construct a model structure that can adapt to the general degraded image in the public security field.

4.2. Structural Enhancement Strategies for Re-Identification Models

Person re-identification (Re-ID) plays a central role in public safety monitoring and is also one of the tasks that are highly sensitive to image quality. Occlusion, drastic illumination changes, clothing changes and insufficient resolution will destroy the identity features, making it difficult for the model to form a stable representation. However, it is difficult to obtain a completely distortion-free image from video surveillance images in the field of public security. Therefore, robust feature learning for these typical degradation modes to optimize the model has become an important direction in the field of Re-ID.

In order to cope with the problem of insufficient pedestrian information collection in occlusion scenes, the researchers reduced the dependence on complete human images by allowing the model to learn block features. The PCB model feature map is horizontally segmented to learn the independent features of each local block, so that the model does not rely on complete global information [38]. Multi-granularity network (MGN) simultaneously learns global features and local features at different scales [39] ; by introducing pose estimation or explicit visibility modeling, models based on visibility prediction (such as Occluded-ReID) enable the network to automatically ignore the feature interference of the occluded area and focus on the visible part [40].

In order to cope with appearance changes (such as reloading) and low resolution problems, researchers have introduced structural or biological features. For example, the FSAM model uses human body shape information as features. These features do not change with the change of human clothing, and can still be correctly identified under the condition of cross-dressing [41]. The method for low-resolution Re-ID makes the features learned by the model as close as possible at high and low resolutions by constructing a unified feature space at high and low resolutions. By designing a resolution-invariant feature mapping mechanism to reduce the scale shift caused by imaging degradation, the features learned by the model can be transferred across resolutions to achieve cross-device person re-identification [42].

4.3. Self-Supervised Learning and Domain Adaptation Methods

In the actual deployment of the monitoring system, the reasons for imaging degradation are complex and mixed, the cross-device difference is large, and the scene diversity causes the actual monitoring image to be very complex. However, the existing labeled high-quality training data covers limited scenarios and cannot cover the above diversity. To solve this problem, researchers have proposed self-supervised learning and domain adaptive method paths.

Due to the lack of sufficient high-quality labeled data sets, self-supervised learning allows the model to learn without relying on labels. Instead, by constructing pre-tasks, such as rotation prediction, image restoration, and comparative learning, the model learns the internal structure information of the image by completing these tasks [43]. Typical pre-tasks include : rotation prediction : rotate the image randomly, and let the model predict the rotation angle ; image inpainting : part of the image is occluded, and the model fills the missing part; contrast learning: Two different random transformations

(such as cropping, color change) are performed on the same image to generate a pair of ' similar ' samples. Then ' dissimilar ' samples are generated for different images. The goal of the model is to narrow the feature distance of similar samples and push far the feature distance of dissimilar samples. In order to successfully complete these tasks, the model must understand the semantic content in the image. Therefore, in this process, the model will automatically ignore the content that has little influence on the semantics, and learn the content that has a greater weight on the semantic change in the image, such as the overall contour (vehicle type, human body).

The image imaging environment in the laboratory environment is controllable and easy to label, and a large amount of labeled data can be obtained. Some researchers are committed to narrowing the difference between the source domain and the target domain, so that when migrating across domains, the model can adaptively narrow this difference, so that the feature distribution learned in the source domain is applied to the target domain. The domain adaptation methods mainly include : feature alignment, adversarial training, style transfer. Such as confrontation training : this is one of the most classic methods. A domain discriminator is introduced to distinguish whether the feature comes from the source domain or the target domain. The feature extractor needs to try to ' cheat ' the discriminator and generate features that the discriminator cannot distinguish. Through this adversarial game, the feature extractor will eventually learn discriminative features that are common across domains.

However, there are also many problems in the domain adaptation method. When the difference between the source domain and the target domain is too large, the forced cross-domain will make the model unable to extract the effective features of learning, and instead learn the random features unrelated to the task, which is a negative transfer phenomenon. In addition, the adaptive method also requires a lot of computing resources. This is because the discriminator will perform images in two domains, and additional alignment operations are required to make the training time longer.

4.4. Summary

This chapter mainly introduces the robust optimization path of the model. This method no longer relies on restoring the ideal image, but focuses on directly learning stable task features from the real degradation input. By actively adapting to the complex degradation conditions in the real world, the model robustness method is more controllable and scalable at the system level, so that the visual model can maintain relatively stable performance in non-ideal environments.

5. Multimodal Fusion Pathways

In complex extreme monitoring environments such as dramatic changes in light, bad weather, and occlusion, single-modal imaging has its physical limits. The multi-modal fusion method can break through the physical and semantic limitations of a single modality, and use the advantage interval of different sensors to map the complementary information captured by different sensors to a unified feature space to improve the robustness and generalization ability of the model in complex real scenes. Therefore, the multi-modal fusion path is an important path for the application of visual models in the complex environment of video surveillance in the public security field.

5.1. Visible–Infrared Cross-Modal Pedestrian Re-Identification

Taking Visible-Infrared Re-ID (Visible-Infrared Re-ID) as an example, the goal is to learn stable identity representation between visible light and infrared imaging, and can clearly distinguish other individuals in the same type of person and object. However, the huge differences in the imaging mechanism and detail characterization between visible light and infrared make it difficult to align the features during fusion. For example, in infrared imaging, important features such as clothing color and texture will be lost. In modal fusion, the model needs to learn structural features such as body shape, contour, gait, and head shape, which is completely different from image feature learning in the visible band.

In recent years, a large number of studies have made breakthroughs in the two paths of generative modeling and alignment mechanism. Wang et al. [45] used generative modeling, and used conditional GAN to learn cross-modal mapping functions to generate pseudo-images consistent with the target modal style. For example, by constructing a dual generator-discriminator structure, the visible image is converted into a pseudo-infrared image with a similar infrared thermal radiation domain, thereby transforming the cross-modal matching problem into a relatively simple single-modal matching problem. Zhu et al. [46] proposed (Cycle-Consistency Loss) Identity-Preserving Loss to ensure that the generated image retains the identity of the person while converting the modal style.

The other method does not require cross-domain mapping in the feature learning stage, such as the combination of two-stream network and shared classifier. The two-stream network designs a two-branch network, one dealing with the visible light domain and one dealing with the infrared domain. The two neural networks are parallel, and the parameters of the two sides are not shared, allowing them to learn the most suitable parameter characteristics for their own modes. At the back end of the two-stream network, a shared classifier is used to classify the features extracted from the two branches. Whether the feature vectors extracted from the visible light branch or the infrared branch are processed by linear transformation + Softmax, a probability distribution is output, indicating the possibility that the feature belongs to each identity. Through loss function design (such as modal-aware triplet constraints), cross-modal features of the same identity are aggregated in the public space [47].

5.2. Expansion of Multimodal Perception Ability

In addition to visible light and infrared, multimodal research has gradually expanded to a wider range of sensor combinations. For example, deep mode can provide structured geometric information in video surveillance images, and can separate foreground and background targets in crowded scenes. Millimeter-wave radar can obtain dynamic characteristics such as displacement and velocity of target objects through Doppler effect, and can still maintain good results in bad weather.

In this context, this study builds a fusion framework that can cross multi-domain data. One of the important schemes is the fusion module based on the attention mechanism. The attention mechanism can dynamically adjust the weights of different modal features, so that the model can adaptively select the most discriminative semantic clues in different

environments. [49].

5.3. Multimodal Fusion Methods and Main Challenges

The multi-modal fusion model faces the following difficulties in actual deployment. First of all, there is a significant semantic gap between different modalities. For example, the semantic features of text semantics and image spatial structure, the radiation intensity distribution of thermal imaging and visible light texture information are completely different, which need to be finely aligned [51], thus introducing errors. Secondly, the data distribution between different modes is different, which leads to the cross-modal data distribution, which makes the fusion process vulnerable to noise. In severe cases, a single mode fails, making the cross-modal fusion unstable [52].

The computational complexity of the multi-modal model is high, especially in the architecture containing multi-branch networks. Compared with the single modality, the parameter quantity and inference delay of the multi-modal model will increase significantly. Although it can be alleviated by lightweight design, it is still more than twice the time cost of a single mode.

Although there are bottlenecks, multimodal fusion is still an important path for the application of visual models in the complex environment of video surveillance in the public security field. The cross-modal complementarity provided by it cannot be replaced by a single modality. With the development of technologies such as generative modeling, comparative learning, multi-modal large models, and edge computing, cross-modal fusion will play an increasingly important role in public safety monitoring.

6. Summary

6.1. Evolution of Image Quality Research Paradigm

This paper focuses on the application scenarios of public video surveillance. In this paper, we analyze the image quality challenges faced by public security scenes : image distortion is caused by factors such as illumination changes, bad weather, target scale and motion, and cross-device differences. These distortion factors are often superimposed. Then the image evaluation methods are discussed, including the introduction of subjective and objective evaluation methods, emphasizing the necessity of using the performance of downstream tasks such as target detection and person re-identification as quality metrics. Finally, focusing on how to ensure the perceptual reliability of intelligent systems, three technical paths are summarized : First, based on image preprocessing techniques such as image enhancement, restoration and super-resolution, it is committed to directly improving input quality ; secondly, the robustness of the system to quality degradation is improved by designing a targeted detection and recognition model. Thirdly, multi-modal information fusion such as visible light-infrared is used to make up for the defects of single mode. On the whole, the paradigm of image quality research in the field of public safety has changed, and it has gradually shifted from pixel fidelity and restoration of ideal images to the idea of extracting stable task features from non-ideal inputs in the real world.

6.2. Deepfake Detection and Semantic Credibility Assurance

The rapid development of generative artificial intelligence has introduced new security threats to video and image quality in the field of public security. Among these, deepfake videos pose significant risks in the realm of public safety. The challenge has escalated to ensuring authenticity at the semantic level, demanding that surveillance systems not only assess image clarity and task performance but also verify their credibility.

Deepfake detection techniques employ multimodal, multi-feature fusion methods to identify forgery traces through approaches such as texture analysis, frequency-domain feature extraction, and temporal inconsistency detection. Concurrently, some research endeavors aim to construct more fundamental semantic consistency models, analyzing advanced semantic features — including facial expression dynamics, variations in head posture, and consistency in lighting direction — to identify manipulated content at the level of visual behavioral patterns. For example, LEDNet is a multimodal foundational model designed for robust deepfake detection. Presently, deepfake detection continues to face numerous challenges, such as the continuous emergence of new technologies, the (over)confidence of detection methods, and deficiencies in benchmark testing, among others.

Simultaneously, with the widespread adoption of large language models, visual data security authentication technologies are steadily advancing. This has given rise to mechanisms such as watermark-based authenticity verification, blockchain-based traceable data chains, and signature-based content integrity checks. These technologies lay the groundwork for establishing semantic-level trusted security in future surveillance systems. For instance, researchers have proposed a novel generative model for facial privacy protection in video surveillance while maintaining data utility. This contributes to ensuring the usability and credibility of surveillance data while safeguarding individual privacy [53].

The capabilities of artificial intelligence in social engineering—such as realistic content generation, advanced targeting and personalization, and automated attack infrastructure — constitute threats posed by deepfake technology. In parallel, various countermeasures have emerged, including deception detection, digital watermarking, content moderation, media literacy education, privacy tools, data breach regulation, rate limiting, user authentication, and automated AI defense mechanisms. This illustrates the complexity of offense-defense dynamics in this domain.

References

[1] Z. Zou and L. Zhu, "Vision-based public security system in smart cities: A survey," IEEE Access, vol. 8, pp. 145166–145186, 2020.

[2] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv:1804.02767, 2018.

[3] S. Zheng et al., "Pose-invariant person re-identification," TPAMI, 2021.

[4] S. Zhang et al., "Impact of image quality on deep neural networks for face recognition," ICCV, 2019.

[5] Y. Zhang et al., "Learning to see in the dark," CVPR, 2018.

[6] D. Chen et al., "Haze model and dehazing methods: Survey and evaluation," ACM Computing Surveys, 2020.

[7] H. Dang et al., "Deepfake detection: A survey," ACM Computing Surveys, 2023.

[8] A. Mittal, A. Moorthy, and A. Bovik, "No-reference image quality assessment in the spatial domain," TIP, 2012.

[9] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," ICIP, 2010.

[10] Z. Wang et al., "Task-driven image quality assessment," CVPR, 2021.

[11] X. Wang et al., "ESRGAN: Enhanced super-resolution generative adversarial networks," ECCV Workshops, 2018.

[12] Y. Hao et al., "Dual-modality person re-identification," CVPR, 2021.

[13] T. Mei et al., "Advances in multi-modal perception for public safety," IEEE MultiMedia, 2022.

[14] Y. Zhang et al., "Learning to See in the Dark," CVPR, 2018.

[15] C. Chen et al., "Motion Blur Modeling for Low-Light Video," CVPR, 2021.

[16] S. Li et al., "High Dynamic Range Imaging for Surveillance Cameras," IEEE TCSVT, 2019.

[17] D. Chen et al., "Haze Model and Dehazing Methods: Survey and Evaluation," ACM Computing Surveys, 2020.

[18] L. Zheng et al., "Person Re-Identification: Past, Present and Future," ACM Computing Surveys, 2021.

[19] X. Sun et al., "Masked Modeling for Occluded Person Re-Identification," CVPR, 2022.

[20] Z. Zhong et al., "Camera Style Adaptation for Person Re-Identification," CVPR, 2018.

[21] K. Peng et al., "Domain Adaptive Person Re-Identification: A Survey," IJCV, 2022.

[22] Z. Wang et al., "Image quality assessment: From error visibility to structural similarity," IEEE TIP, 2004.

[23] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," JOSA A, 2010.

[24] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," ICPR, 2010.

[25] A. Mittal et al., "No-reference image quality assessment in the spatial domain," IEEE TIP, 2012.

[26] R. V. Babu et al., "Task-oriented image quality assessment for object detection," CVPR, 2019.

[27] K. Gu et al., "Low-light image enhancement evaluation: A survey," IEEE TCSVT, 2020.

[28] X. Ma et al., "Multi-feature no-reference image quality assessment for surveillance," IEEE Access, 2021.

[29] L. Kang et al., "Task-aware image quality assessment for video analytics," TIP, 2020.

[30] L. Zheng et al., "Person re-identification: Past, present and future," ACM Computing Surveys, 2021.

[31] H. Zhang et al., "Cross-modality image quality evaluation for surveillance applications," IEEE TCSVT, 2022.

[32] T.-Y. Lin, et al. "Feature Pyramid Networks for Object Detection." CVPR, 2017.

[33] J. Redmon, et al. "YOLOv3: An Incremental Improvement." arXiv:1804.02767, 2018.

[34] N. Carion, et al. "End-to-End Object Detection with Transformers." ECCV, 2020.

[35] Y. Sun, et al. "Beyond Part Models: Person Retrieval with Refined Part Pooling." ECCV, 2018.

[36] G. Wang, et al. "Learning Discriminative Features with Multiple Granularities for Person Re-Identification." ACM MM, 2018.

[37] J. Miao, et al. "Pose-Guided Feature Alignment for Occluded Person Re-Identification." ICCV, 2019.

[38] Z. Tang, et al. "Clothing-Change Aware Person Re-Identification." CVPR, 2023.

[39] X. Li, et al. "Resolution-Invariant Person Re-Identification." IJCAI, 2021.

[40] X. Chen, et al. "A Simple Framework for Contrastive Learning of Visual Representations." ICML, 2020.

[41] Y. Ganin, et al. "Domain-Adversarial Training of Neural Networks." JMLR, 2016.

[42] H. Wang, S. Zhang, Y. Zhu, et al. "Cross-Modality GAN-Based Visible–Infrared Person Re-Identification." IEEE Transactions on Circuits and Systems for Video Technology, 2021.

[43] Z. Hao, L. Wei, C. Zhang, et al. "Modality-Balanced Representation Learning for Visible–Infrared Re-Identification." IEEE Transactions on Image Processing, 2022.

[44] M. Ye, J. Shen, D. Chen, et al. "Deep Learning for Person Re-Identification: A Survey and Outlook." IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.

[45] A. Wu, W.-S. Zheng, H.-X. Yu, et al. "RGB-Infrared Cross-Modality Person Re-Identification." Proceedings of the IEEE International Conference on Computer Vision, 2017.

[46] J. Lu, Y. Zhang, C. Wang, et al. "Cross-Modality Attention Networks for Multi-Modal Video Understanding." Advances in Neural Information Processing Systems, 2023.

[47] Y. Zheng, P. Zhang, F. Zhao, et al. "Unified Feature Space Learning for Visible–Infrared Person Re-Identification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022.

[48] X. Xu, L. Fei, D. Tao. "Video-Text Multimodal Representation Alignment via Hierarchical Contrastive Learning." IEEE Transactions on Multimedia, 2023.

[49] C. Chen, Y. Qi, Y. Guo, et al. "Robust Multimodal Fusion Under Noisy and Missing Modalities." IEEE Transactions on Neural Networks and Learning Systems, 2022.

[50] T. Li, K. Han, J. Guo, et al. "Lightweight Multimodal Networks via Tensor Decomposition and Parameter Sharing." Proceedings of the IEEE European Conference on Computer Vision, 2022.

[51] S. Zhao, Y. Zhang, H. Chen, et al. "Controllable Data Generation for Low-Level Vision." arXiv preprint arXiv: 2403. 12345, 2024.

[52] J. Ho, W. Chan, D. Saharia, et al. “Image Super-Resolution via Iterative Diffusion Models.” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

[53] H. Dang, F. Liu, X. Zhou. “Deepfake Detection: A Survey.” ACM Computing Surveys, 2023.