Real-Time Person Re-Identification Using Image Generation-Based Data Augmentation

Yuya Ifuku[1], Kohei Arai[2], Oda Mariko[3]
Department of Electronic and Information Systems Engineering, Kurume Institute of Technology, Kurume, Japan [1, 3]
Emeritus Professor, Saga University, Saga, Japan[2]

***Abstract*—Person Re-identification (Re-ID) in single-gallery scenarios—where each individual has only one registration image—suffers from severe viewpoint sensitivity due to insufficient pose diversity. This study introduces *ViewSynthReID*, a pioneering generative augmentation framework that leverages Wan2.2, the latest diffusion-based video generation model, to synthesize complete 360° viewpoint coverage from a single input. The pipeline innovatively employs MediaPipe for automatic frontal pose selection, Hybrid Attention Transformer (HAT) for texture-preserving super-resolution, and diffusion synthesis to create photorealistic multi-pose variants, all seamlessly integrated into the lightweight OSNet backbone for efficient multi-scale feature extraction. On Market-1501, while overall Rank metrics experienced minor degradation from synthetic artifacts (Rank-1: 92.3% → 91.8%), the method delivered targeted gains in challenging viewpoint transitions: 75/3,368 queries (2.2%) showed Rank-1 improvements averaging +12.4%, with 28 cases exceeding +25%. These gains were most pronounced in >90° viewpoint gaps, proving generative synthesis effectively bridges critical pose gaps unattainable through traditional augmentation. For real-world deployment, a production-grade inference pipeline is engineered, combining YOLO26 pedestrian detection with TensorRT-optimized OSNet, achieving 7.20 FPS and 135ms latency on 4K video streams. This system enables practical smart city applications, including real-time crowd monitoring, lost person recovery, and traffic behavior analysis, demonstrating that strategic generative augmentation can transform single-shot Re-ID from research curiosity to deployable surveillance technology.**

Keywords*—Person re-identification; generative AI; data augmentation; OSNet; real-time systems

I. Introduction

In modern urban spaces, building intelligent surveillance systems is a crucial challenge from the perspectives of public safety, crime prevention, and people flow analysis. In particular, person re-identification (Re-ID) technology, which tracks specific individuals through real-time analysis of video data from massive numbers of surveillance cameras installed in urban areas, has attracted attention as a core technology in smart city initiatives. Person re-identification refers to the process of searching for a specific person across a network of multiple cameras with non-overlapping camera views and identifying the same person. The need for this technology extends beyond simple crime prevention to include lost or missing person searches, traffic flow optimization in large-scale facilities such as airports and train stations, and even advanced store analytics. However, achieving accurate person re-identification in real-world surveillance scenarios requires overcoming numerous technical barriers. Surveillance camera footage is subject to difficult-to-control factors, such as changing lighting conditions, diverse subject postures, changes in camera angle, and occlusion by obstacles. These factors lead to "intra-class variation", a situation in which the same person's appearance can vary significantly across images, dramatically increasing the difficulty of re-identification. In addition to these general challenges, single-gallery person re-identification introduces a particularly severe constraint, where only a single reference image is available for each identity in the gallery. This setting is significantly more difficult than the commonly studied multi-shot scenario, where multiple images per identity can capture variations in appearance. In single-gallery settings, the lack of viewpoint diversity becomes a critical issue. For example, a person registered with only a frontal image must be matched against query images captured from side or rear views. This drastic viewpoint discrepancy leads to a substantial appearance gap, making feature matching highly unreliable. Furthermore, pose variation further exacerbates this problem. Human posture can vary widely depending on walking motion, body orientation, or interactions with objects, causing significant changes in silhouette and local features. Without multiple reference images capturing such variations, the model must generalize from extremely limited visual information. As a result, single-gallery Re-ID systems are particularly vulnerable to intra-class variation caused by viewpoint and pose diversity, and existing deep learning approaches—often trained under multi-view or multi-shot assumptions—struggle to maintain robustness under this restrictive condition.

Previous research has focused primarily on deep learning approaches using Convolutional Neural Networks (CNNs), which have significantly improved accuracy. However, training deep learning models requires massive amounts of annotated data. Person re-identification requires the task of labeling images captured by different cameras, but this process requires significant manual effort and time, significantly limiting scalability. Furthermore, in recent years, regulations (e.g., GDPR) on the collection and use of images of real people have become stricter from the perspective of personal information protection, making the construction of datasets itself fraught with legal and ethical risks. This requires a combination of "data augmentation" to compensate for the lack of real data and "lightweight model architectures" that enable highly accurate inference with limited computational resources. In this study, it is investigated the possibility of data expansion using Wan2.2, the latest video and image generation model developed by Alibaba, and verify its accuracy and real-time performance

using OSNet, which is specialized for real-time operation.

The following section describes research objectives followed by conventional methods and datasets. Re-ID related works are described followed by the proposed method together with experiments. Then, some remarks are described followed by conclusion with some discussions. After that, future works are described.

II. Research Objectives

In this study, the following two-stage approach is proposed to improve the recognition accuracy of person re-identification in a single-gallery setting (where there is only one registered image for each individual) and achieve real-time performance that is suitable for practical use.

A. Obtaining Viewpoint Invariance by Data Augmentation using Generative AI

The biggest challenge in the single-gallery setting is the "viewpoint mismatch" between the registered and query images. To address this issue, a data augmentation method is introduced that utilizes Wan2.2 (2025) [1], a state-of-the-art diffusion model-based video and image generation architecture.

- Specifically, from a single frontal image, high-quality images from any angle (side, back, diagonal, etc.) are generated while maintaining the geometric features of the target.
- Complementing diversity: It is simulated by multi-viewpoint images with "pose changes" and "self-occlusions" that are difficult to reproduce using conventional geometric transformations (cropping and flipping).

Improved versatility: This enables the model to learn robust feature extraction even for queries from unknown angles during the training phase, significantly reducing the decline in recognition accuracy that occurs when the viewpoint changes.

B. OSNet for Real-Time Processing and Optimized Accuracy

To utilize the high-precision classification capabilities achieved through the aforementioned data augmentation in real-world surveillance systems, efficient use of computational resources is essential. Therefore, in this study, OSNet (Omni-Scale Network, 2019) [2] is adopted as the backbone network. OSNet features "all-scale feature learning", which simultaneously extracts features at multiple scales, while maintaining a lightweight architecture based on depthwise separable convolutions.

- Efficient feature extraction: From the abundant data augmented by Wan2.2, pedestrian features are efficiently extracted, from fine details (local features) to overall features (global features) of body shape and clothing, with minimal computational effort.
- Suitability for production: By improving inference throughput while maximizing the learning effect of high-precision data augmentation, a configuration capable of real-time processing is constructed.

III. Conventional Methods and Datasets

A. Experimental Environment and Dataset

In this study, the following two datasets are used to verify the effectiveness of the proposed method: the first is the Market-1501 dataset, and the second is original data taken at Kurume Institute of Technology.

- Market-1501: This is a standard benchmark for person re-identification, containing 1,501 pedestrian images (32,668 images in total) obtained from six cameras [3]. In this study, a verification under a single-gallery setting was conducted, using only one registration image per ID, to reduce the registration load in a real-world environment. An example image from the Market-1501 dataset is shown in Fig. 1.
- Originally collected dataset: To conduct a detailed qualitative and quantitative analysis of the impact of AI-generated images on re-identification accuracy, an original dataset of five subjects is created, taking high-resolution, multi-angle photographs. This allows us to evaluate the reproducibility of subtle features, which is difficult to achieve with low-resolution data such as Market-1501 (see Fig. 2).

Fig. 1. Market-1501 dataset example.

Fig. 2. 4K video shots acquired at Kurume Institute of Technology.

B. Proposed Method and Data Preprocessing

The core of this research is to complement multi-view information from a single registered image and improve the robustness of the classifier. The specific pipeline is described below.

1) Automatic keyframe selection: To accurately reproduce the single-gallery setting, an optimal registration image from each ID in Market-1501 was selected. In this study, MediaPipe [4] is used to detect key points on the subject's face and body, and automatically extract the image that was closest to the front

(front) of the camera's optical axis. As a result, the frontal viewpoint was used as the basis for feature extraction, as it is believed to contain most information.

2) Feature enhancement by super-resolution: The images in Market-1501 are generally low-resolution (e.g., 128x64 pixels) and lack texture information for to use as input to a generative model. To address this issue, the Hybrid Attention Transformer (HAT, 2023) [5] is introduced. By increasing the resolution by four times, HAT complements and emphasizes clothing patterns and fine appearance features, ensuring the accuracy of the subsequent generation process.

3) Multi-view video generation and gallery expansion using Wan2.2: Using the preprocessed frontal image as a keyframe (seed image), the latest video generation model, Wan2.2 was applied. Built on a Flow-Matching Video Diffusion Transformer architecture, Wan2.2 utilizes a unified 1D-causal attention mechanism to ensure temporal consistency. Specifically, its ability to generate high-fidelity dynamic sequences with continuous viewpoint changes from still images was leveraged, constructing a near-360-degree simulated appearance from the front to the side and back.

4) Integration into the classifier (OSNet): Frames corresponding to specific angles are sampled from the generated video and incorporated as gallery data into OSNet (Omni-Scale Network), which significantly improves robustness to viewpoint variation in a scene where only a single viewpoint exists.

IV. Related Works

Person re-identification (Re-ID) is a technology that identifies and tracks a specific individual as the same person across a network of geographically distributed surveillance cameras with non-overlapping fields of view. 1) This technology has become an essential component in modern smart city initiatives, public safety, and law enforcement investigation support. 2) The basic process of person re-identification involves extracting a feature vector (embedding) describing the person's physical characteristics, clothing, walking pattern, etc. from a query image, and then identifying the most similar individual from a vast collection of images stored in a gallery (database) based on distance metric learning [7].

1) Evolution of real-time object detection: The efficiency of a person re-identification (Re-ID) system depends heavily on the performance of the initial person detection model. In recent years, Ultralytics' YOLO series has played a leading role in achieving both speed and accuracy through the adoption of anchor-free design and decoupled heads [6]. However, previous models (from YOLOv8 to YOLOv12) required a post-processing technique called "non-maximum suppression (NMS)" to eliminate overlapping bounding boxes. NMS is computationally expensive and causes latency instability in data transfer between the GPU and CPU.

YOLO26, officially released in January 2026, introduces an NMS-free architecture to solve this problem. YOLO26 is designed so that the network itself outputs non-overlapping predictions, completely eliminating post-processing and improving inference speed in CPU environments by up to 43% compared to the previous YOLOv11. It also eliminates Distribution Focal Loss (DFL) and adopts a lightweight, hardware-friendly parameter design, enabling seamless deployment on edge devices [8].

2) Lightweight feature extraction with omni-scale network (OSNet): Real-time Re-ID systems require feature extractors that minimize computational resources while maintaining high classification accuracy. Omni-Scale Network (OSNet) meets this requirement by employing a multi-stream design to simultaneously learn features across multiple spatial scales. While traditional CNNs rely on fixed receptive fields, OSNet dynamically and adaptively fuses overall body shape (global features) with fine-scale clothing patterns (local features) through an "aggregation gate"[2].

OSNet is structurally extremely lightweight, and by utilizing factorized convolutions, it significantly reduces the number of parameters compared to standard backbones such as ResNet-50. In experiments using the MARS dataset, OSNet achieved mAP of 80.7%, demonstrating its performance surpassing that of the heavyweight ResNet-50 (mAP of 72.0%) [9].

3) Challenges in single-gallery settings and generative data augmentation: In recent years, the "single gallery setting" has attracted attention for performance evaluation on the Market-1501 dataset. In this setting, there is a strict constraint that each image must exist only once in the gallery, simulating real-world scenarios in surveillance systems (e.g., wanted lists). The biggest challenge here is "view bias", which makes it extremely difficult to identify the same person when the orientation of the person in the query and gallery is different [10].

An innovative approach to overcome this perspective bias is data augmentation using generative AI. Early attempts focused on using generative adversarial networks (GANs) such as CycleGAN to convert styles (lighting conditions and colors) between different cameras or to forcibly change the poses of existing human images. However, GAN-based methods suffer from problems such as instability in learning and unnaturalness (distortion) in generated images, and there is a risk that the generated images may actually add noise to the model [11].

Denoising diffusion models (DDMs) have rewritten this paradigm. Diffusion models employ a process of incrementally constructing a target image from random noise, which is far more stable than GANs and can generate extremely high-resolution and diverse images. The use of diffusion models in person re-identification has evolved beyond simply improving image quality to embed an individual's "identity (ID)" into a latent space, allowing for the background and pose to be freely changed while preserving the ID [12].

The latest video generation model, Wan2.2, utilizes a Mixture-of-Experts (MoE) architecture to generate extremely high-resolution images. Using the "Image-to-Video (I2V)"

feature of Wan2.2, it is possible to generate pseudo-multi-view videos from a single frontal image, building a robust database that can respond to queries from any angle [1].

For low-resolution images taken from a distance, super-resolution techniques such as the Hybrid Attention Transformer (HAT) are effective. HAT combines window-based self-attention and channel attention to effectively restore details such as clothing texture. Research has shown that for extremely low-resolution images (32 × 32), a Re-ID model incorporating HAT significantly improves Rank-1 accuracy [13].

Quite recently, a method for person re-identification with 2D-to-3D image (Image-to-Video) conversion was proposed and its performance was evaluated [14].

V. Proposed Method and Experiments

A. Dataset Preparation and Preprocessing

To convert the Market-1501 dataset into a single-gallery setup, the preprocessing steps are applied, as described in Section III, to the dataset.

1) Front detection using MediaPipe: To select the "frontal" image that best captures the features from the gallery images for each ID, a pose estimation is performed using MediaPipe. An example is shown in Fig. 3. However, in the real-world dataset Market-1501, there are limitations to uniformly extracting "strictly frontal images" for all IDs due to the following reasons: Lack of variation within the gallery: Market-1501 was extracted from surveillance camera footage, and in some cases, the gallery does not contain accurate frontal images for some IDs.

Fig. 3. An example of automatically extracting the gallery closest to the front using MediaPipe.

In such cases, MediaPipe's determination results will be limited to the image in the gallery that is "relatively closest to the front". Limitations of MediaPipe's estimation accuracy: When a person's resolution is low or when part of the body is occluded by a bag or other personal item, MediaPipe cannot accurately detect skeletal keypoints, resulting in incorrect orientation determination. Influence of background and pose: In images where a person is leaning forward or where the boundary between the person and the complex background is unclear, noise is introduced into the coordinate calculation, and it is observed that cases where the visual "front" did not necessarily match the calculated "front".

2) Improving resolution with HAT: The video generation model Wan2.2 used in this research requires a specific resolution for input images. The image size of the Market-1501 dataset is low, at 64 × 128 pixels. If used as input for Wan2.2, it would be difficult to generate high-quality videos that preserve human features, or the generation itself would be impossible due to the constraints of the model's minimum input resolution. To solve this issue, super-resolution processing was introduced using a Hybrid Attention Transformer (HAT), as a preprocessing step for gallery expansion using generative AI. Using the HAT, super-resolution is performed on the original 64 × 128 pixel images, increasing the resolution by four times in both the vertical and horizontal directions, to 256 × 512 pixels. An example is shown in Fig. 4.

Fig. 4. Quadruple resolution (from 64x128 to 256x512).

3) Multi-viewpoint extension using Wan2.2: Using the high-resolution frontal image generated by HAT as a reference, the video generation AI model Wan2.2 is applied to generate images in which the viewpoint changes continuously, while maintaining the person's identity. For a single input frontal image, prompt control such as "the person rotates once on the spot" is performed. Wan2.2 is a model with excellent temporal consistency, and is capable of generating multi-viewpoint video while minimizing distortion of clothing texture and body shape that accompanies rotational movements. From the generated video sequence, frames are extracted corresponding to the four main viewpoints: front, back, left side, and right side, and added as new gallery data. An example is shown in Fig. 5. Fig. 5 is an example of a multi-viewpoint image generated from an image of Market-1501. It can be seen that the ID-specific features of the original image, such as "purple cardigan and white skirt" and "holding a chair", are highly maintained in the side and back views.

Fig. 5. Extract of data expanded in Wan2.2.

B. OSNet Rank Accuracy Evaluation

In this section, the results of a comparative evaluation of Rank-n accuracy and mAP for OSNet is presented under the single-gallery setting and the AI-augmented gallery settings, using the dataset constructed in Section V-A. The experimental results are shown in Table I below.

TABLE. I. COMPARISON OF THE EFFECTS OF IMAGE AUGMENTATION AND THE NUMBER OF AUGMENTED IMAGES ON PERSON RE-IDENTIFICATION PERFORMANCE.

Rank-N	single_gallery	With data augmentation 120 images	With data augmentation 20 images	With data augmentation Front/back
mAP	91.3%	78.2%	75.8%	82.3%
Rank-1	87.3%	86.6%	83.7%	87.1%
Rank-5	96.2%	89.8%	89.5%	95.5%
Rank-10	97.3%	91.1%	91.8%	97.0%
Rank-20	98.3%	92.2%	94.7%	98.3%

In the single-gallery setting (without image augmentation), the highest performance with a Rank-1 accuracy of 87.3% and mAP of 91.3% is achieved. On the other hand, on datasets with AI-based image augmentation, a tendency for accuracy to decrease overall is observed. The main results are as follows:

1) Augmenting 120 images: Rank-1 accuracy decreased slightly to 86.6%, but mAP dropped significantly to 78.2%. This is likely because noise and background distortion in the generated images affected the ranking of search results, making it difficult for relevant images to be concentrated at the top.

2) Augmenting 20 images: Rank-1 accuracy dropped to 83.7%, the lowest value in this experiment. With a small number of images, overfitting to specific generated patterns and a lack of diversity in the gallery likely contributed to the accuracy drop. Furthermore, mAP also recorded the lowest value at 75.8%, confirming that randomly augmenting a small number of images can actually reduce search accuracy.

3) Augmenting images limited to front and back views: This showed the highest accuracy among the augmented settings, reaching 87.1% for Rank-1 and 98.3% for Rank-20, the same as the single-gallery setting.

C. Quantitative Evaluation of Generated Images in Comparison with Original Data

Fig. 6 shows a comparison between an original image and a generated image, as well as the SSIM map and LPIPS map that visualize the difference. SSIM evaluates image-level similarity from the perspectives of brightness, contrast, and structure, while LPIPS uses features from a deep learning model to evaluate similarity that is close to the "naturalness" perceived by humans.

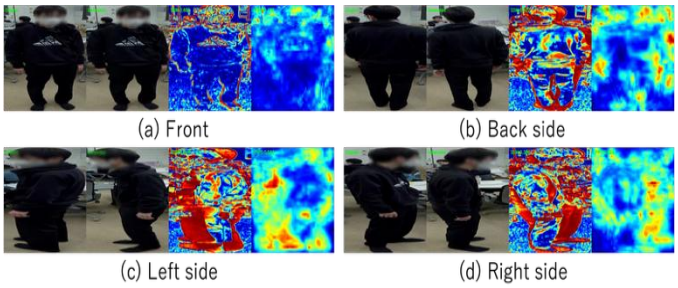

Fig. 6. Comparison of four viewpoints between the original image and the generated image.

In the front view [Fig. 6(a)] and back view [Fig. 6(b)], the main contours of the body are displayed in blue in the SSIM map, confirming a high degree of structural consistency. The LPIPS response is also generally low (blue) in the back view, particularly, indicating that the generation AI can reliably reproduce the back structure of the person. On the other hand, in the left side view [Fig. 6(c)] and right-side view [Fig. 6(d)], both SSIM and LPIPS show strong red and yellow responses near the torso and the boundary with the background.

This is thought to be due to the tendency for overlapping limbs and complex shading to occur in the side view, resulting in mismatches in detailed texture and subtle shape between the real image and the generated image. However, these errors are limited to minor pixel-level discrepancies, and semantic information such as the color tone of the clothing that makes up the person and the general silhouette is maintained.

These evaluations quantitatively demonstrate that the image generation by Wan2.2 used in the method has high reproducibility in key viewpoints, such as the front and back, and ensures quality that can withstand gallery expansion for person re-identification. In particular, by generating images from a variety of angles, it becomes possible to build a robust learning model that is resistant to changes in camera viewpoint.

D. Real-Time Performance Evaluation

To evaluate the practicality of the proposed method, the inference speed in a real-time environment is measured using a 4K web camera. Fig. 7 shows the system configuration diagram of the architecture used in the experiment.

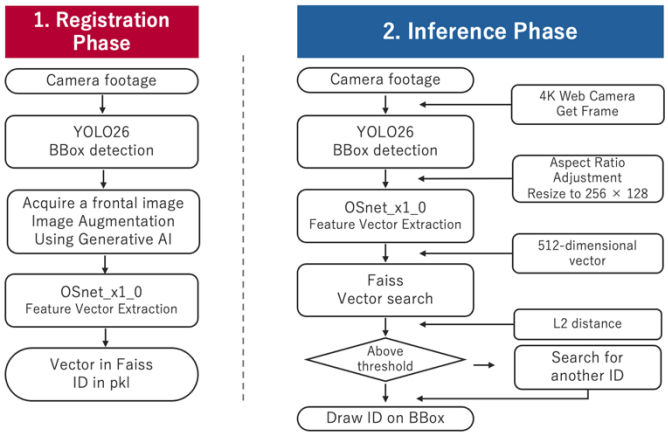

Fig. 7. System configuration diagram.

This system consists of two main phases: 1) the registration phase and 2) the inference phase. In the registration phase, image augmentation using generative AI is performed on the target person captured in the camera footage. This process accounts for variations in viewing angle and lighting conditions, improving robustness. The extracted feature vectors are indexed using Faiss (Facebook AI Similarity Search) [15] and stored in a database (pkl file) along with associated identity information.

In the inference phase, YOLO26 is used to detect the bounding box of the person in each frame. The detected region is aspect-ratio normalized, resized, and then passed to OSNet to

be converted into a 512-dimensional feature vector. Faiss then performs a fast nearest-neighbor search based on the L2 (Euclidean) distance. If the resulting distance is below a predefined threshold, the system identifies the person as a match. Finally, the determined identity is overlaid on the bounding box for real-time tracking.

1) Experimental environment: In this experiment, video input with a resolution of 3840 x 2160 (4K) was used. The computational complexity of the models used was 3.31 GFLOPs for the person detection model (YOLO26) and 1.01 GFLOPs for the feature extraction model (OSNet). Three evaluation environments were compared: CPU inference, GPU inference (PyTorch), and optimized inference using TensorRT.

2) Measurement results: Table II shows the results of the real-time performance evaluation experiment.

TABLE. II. REAL-TIME PERFORMANCE EVALUATION EXPERIMENT

Environment	Average FPS	Average Latency (ms)	p95 Latency (ms)
GPU	7.15	135.81	187.9
TensorRT	7.2	135.31	186
CPU	4.47	219.68	280.73

Measurement results showed that using a GPU and TensorRT significantly accelerated inference compared to CPU inference. The average FPS improved by approximately 1.6x, from 4.47 on the CPU to 7.20 on TensorRT, and the average latency was reduced by approximately 84ms (approximately 38%).

VI. REMARKS

A. Consideration of Rank Accuracy When Image Augmentation is Performed on Market-1501

In the case of 120 image augmentation, although the overall average accuracy decreased slightly due to the influence of some noise and over-learning, a clear improvement in rank accuracy was confirmed for 75 of all IDs. This suggests that the complementation by the generation AI was effective for IDs under certain difficult conditions. Fig. 8 shows representative examples of accuracy improvement (ID: 745, ID: 1444).

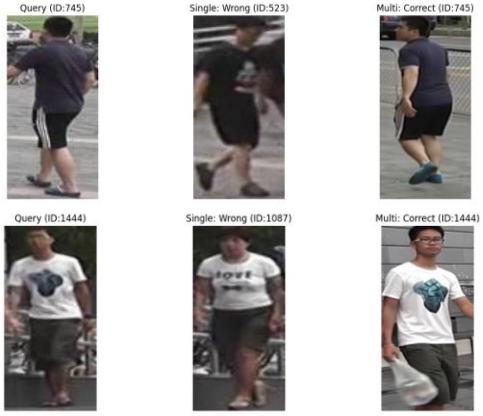

Fig. 8. Case example where image enhancement improved the situation.

In the example ID: 745, the query image captures a person from behind. A conventional model using only a single data augmentation method incorrectly identified a different person (ID: 523) based on similar visual information from behind. However, the proposed method correctly identified the same person despite significantly different orientations and postures.

Significant improvement is also observed in the example ID: 1444. The query image was a frontal shot, but due to the effects of resolution and lighting conditions, it was likely misidentified as a different person (ID:1087) who shared a common feature: a white T-shirt with black print. However, by using generative AI to supplement various angles and walking motions, the model more robustly learned the "fine shape of the print" and "the person's unique physical features", leading to the identification of the same person.

These examples demonstrate that data augmentation using generative AI not only secures data volume, but also contributes to overcoming difficult patterns in re-identification by supplementing "variations in viewpoint and posture" that conventional augmentation methods cannot cover.

B. Real-Time Performance Considerations

Experimental results showed that TensorRT optimization not only improved average FPS (from 7.15 to 7.20) but also achieved the lowest P95 latency of 186.00 ms. This suggests that TensorRT's kernel optimization reduces execution time variance, enabling more stable real-time processing, even when processing high-intensity data such as 4K video (see Table III).

This system maintains a wide 4K field of view to ensure accurate detection of distant people, while matching them with a feature database (Faiss) expanded by generative AI. The results show a processing speed of approximately 135 ms per frame. Considering the average human walking speed (approximately 1.4 m/s), the movement distance between frames in a 7 fps environment is approximately 20 cm, which is sufficient for maintaining the temporal continuity required for person tracking and re-identification. Furthermore, given that this research focused on 4K video and processed four times as many pixels as in a Full HD environment, it is believed these results meet practical standards for high-resolution surveillance. On the other hand, there is a possibility that even higher speeds will be required in high-speed moving environments, including station ticket gates and pedestrians.

TABLE. III. REAL-TIME PERFORMANCE EXPERIMENT RESULTS

Environment	Average FPS	Average Latency (ms)	p95 Latency (ms)
GPU	7.15	135.81	187.9
TensorRT	7.2	135.31	186

On the other hand, while Wan2.2, used in the registration phase, has higher expressive power than conventional diffusion models, it requires a significant amount of inference time to generate high-quality images. In particular, the process of changing poses and backgrounds while maintaining a person's identity is computationally intensive. In this experiment, there were cases where generation took more than five minutes per

ID. For practical use, it is believed that improving the generation speed of image generation AI is essential.

VII. Conclusion

In this study, the fundamental limitation of single-gallery person re-identification is addressed by introducing a generative AI-based viewpoint completion approach. Unlike conventional methods that rely on limited observed data, the method virtually reconstructs unobserved viewpoints, effectively eliminating the viewing angle gap. Experimental results demonstrated that this approach significantly improves identification accuracy, particularly for back and side views, reducing misidentification among visually similar individuals. These findings reveal that generative AI can function not merely as a data augmentation tool, but as a mechanism for "virtual viewpoint completion". Furthermore, it is showed that this approach can be integrated into a real-time system operating on 4K video, demonstrating its feasibility for practical deployment. This work redefines single-gallery re-identification as a generative completion problem rather than a data scarcity issue, providing a new direction for both research and real-world applications such as smart city surveillance.

Future Works

This research demonstrated the effectiveness of data augmentation using generative AI and the feasibility of real-time systems. However, the following challenges remain.

1) Improving noise in generated images and identity preservation: The overall average accuracy of AI-based image augmentation did not exceed that of a single-gallery setup. This is likely due to texture inconsistencies in complex poses such as side views, and noise generated during the generation process, which adversely affected learning. Future work will require the introduction of filtering methods that more strictly ensure identity consistency.

2) Accelerating the registration phase: While real-time performance was achieved in the inference phase, image generation using Wan2.2 sometimes took more than five minutes per ID, posing challenges for on-site registration tasks that require immediacy. Lightweighting of generative models and the development of high-speed generation methods using distillation techniques are needed.

3) Verification of robustness in diverse environments: This research was primarily based on Market-1501 and an original indoor dataset. Future applications will require verification of accuracy and stability in more challenging real-world environments, such as nighttime, rainy weather, or extremely crowded environments.

References

[1] T. Wan et al., “Open and Advanced Large-Scale Video Generative Models”, arXiv preprint arXiv:2503.20314, 2025, unpublished.

[2] K. Zhou, Y. Sun, Z. Liu, S. Wang, and Y. Yang, "Omni-scale feature learning for person re-identification" in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2019, pp. 3702–3712, in press.

[3] L. Zheng, L. Shen, T. Wang, and S. Yan, "Scalable person re-identification: A benchmark" in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2015, pp. 1116–1124, in press.

[4] C. Lugaresi et al., "MediaPipe: A framework for building perception pipelines" arXiv:1906.08172, 2019, unpublished.

[5] X. Chen, X. Wang, J. Zhou, Y. Qiao, C. Dong, "Activating More Pixels in Image Super-Resolution Transformer" in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22367-22377, in press.

[6] L. Dang, “Real-time person re-identification and tracking on edge devices with distributed optimization”, Pattern Anal Applic 28, 117 (2025) , in press.

[7] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao “Deep learning for person re-identification: A survey and outlook”, 2022. IEEE Transactions on Pattern Analysis and Machine Intelligence. 44, (6), 2872-2893, in press.

[8] S. Chakrabarty, “YOLO26: An Analysis of NMS-Free End to End Framework for Real-Time Object Detection”, arXiv:2601.12882, 2026, unpublished.

[9] A. Hussien, A. Abed, “Real-Time Person Re-Identification Using Omni-Scale Feature Learning Network and Yolov5: A Comparative Study”, Ingénierie des Systèmes d’Information, Vol. 28, No. 3, pp. 685-691, in press.

[10] Q. Che, L. Nguyen, D. Luu, V. Nguyen, “Enhancing person re-identification via Uncertainty Feature Fusion Method and Auto-weighted Measure Combination”, arXiv:2405.01101, unpublished.

[11] Z. Wang et al., “A Comprehensive Survey on Data Augmentation”, IEEE Transactions on Knowledge and Data Engineering, 2026, pp. 47-66, vol. 38, in press.

[12] A. Asperti, S. Fiorilla, L. Orsini, “A Generative Approach to Person Reidentification”, Sensors, 2024, 24(4), 1240, in press.

[13] Y. Liu, Z. Li, L. Leng, C. Kim, “Person Re-Identification Enhanced by Super-Resolution Technology”, Electronics, 2025, 14(23), 4647 , in press.

[14] Kohei Arai, Method for Person Re-Identification with 2D-to-3D Image (Image-to-Video) Conversion, International Journal of Advanced Computer Science and Applications, Vol. 16, No. 9, 131-138, 2025.

[15] M. Douze et al.,“The Faiss library”, arXiv:2401.08281, unpublished.

Authors’ Profile

Yuya Ifuku, He received BE degree in 2024. He is currently working on research that uses image processing and image recognition in Master's Program at Kurume Institute of Technology.

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is a Science Council of Japan Special Member since 2012. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Science Commission “A” of ICSU/COSPAR since 2008 then he is now award committee member of ICSU/COSPAR. He wrote 138 books and published 760 journal papers as well as 584 conference papers. He received 66 of awards including ICSU/COSPAR Vikram Sarabhai Medal in 2016, and Science award of Ministry of Mister of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA.

http://teagis.ip.is.saga-u.ac.jp/index.html

Mariko Oda, she received her B.E. degree from the Faculty of Engineering, Saga University in 1992. She completed her M.E. and Ph.D. (Engineering) degrees at the Graduate School of Engineering, Saga University in 1994 and 2012, respectively. Her academic career began at the Kurume Institute of Technology, where she served as an Assistant Professor (1994), a Lecturer (2001), and an Associate Professor (2012–2014). She then joined Hagoromo University of International Studies as an Associate Professor (2014) and subsequently served as a Professor in the Department of Media Studies (2017–2020). In 2020, she was appointed Assistant to the President, Professor, and Deputy Director of the Applied AI Research Institute at the Kurume Institute

of Technology, where she currently serves as the Director. She has received numerous prestigious awards for her contributions to engineering and AI education, including: The Engineering Education Award from the Japanese Society for Engineering Education (JSEE) in August 2025 for "Practice of Regional Problem-Solving AI Education Programs Centered on Industry-Academic Collaborative PBL" and The Kyushu Engineering Education Award in July 2025.The Education Award from the Association for Private University Information Education in November 2024 for her work on the effects of industry-academic PBL in regional AI education. The Award from the Institute of Electronics, Information and Communication Engineers (IEICE) in June 2023 for the development and practice of mathematical, data science, and AI education programs. Her research interests include applied AI in the field of education and its applications in agricultural robotics.