ABSTRACT  Person re-identification (ReID) identifies object IDs in a multicamera environment based on local tracking of city surveillance cameras in public places. This method can improve the performance by learning various features using a convolutional neural network. However, ReID methods are limited in terms of application in practical surveillance environments as the ReID model is trained on public datasets and lacks the ability to generalize images acquired using other cameras. Moreover, although various methods to improve the ReID performance have been proposed, most existing studies did not evaluate ReID performances according to the number of parameters, i.e., ReID models that can be used in a limited memory environment were not considered. In this study, we propose a Tiny Asymmetric Feature Normalized Network, which can be generalized to test at sets acquired in real surveillance environments considering various scale features and can be operated with a limited number of parameters. Moreover, the Gwangju Institute of Science and Technology Practical Person ReID (GPP-reID) dataset, which was used to evaluate the performance of the ReID model, has been distributed and made available to enable applications in real-world surveillance environments. Our proposed method achieved mean average precision (mAP), Rank-1 values of 86.2, 94.7 and 74.8, 85.9 on the Market1501 and Duke Multitracking Multicamera ReID datasets, respectively. In addition, mAP and Rank-1 values of 44.2 and 64.1, respectively, were achieved on the cross-validated, new benchmark dataset, GPP-reID, using a network with one-tenth the number of parameters as the 50-layer residual network.

INDEX TERMS  Convolutional neural network, cross cameras, deep learning, person re-identification, video surveillance.

I. INTRODUCTION  Person re-identification (ReID) is a subtask in image retrieval. It is a technique used for the identification of specific objects acquired in a multicamera network environment. ReID is a method that can match images of a person obtained from different camera views and has recently received considerable attention owing to security problems. A person-type object that must be identified is defined as a query or probe. Person objects that are captured by multiple nonoverlapping surveillance cameras are combined to form a “gallery”. When a query is initiated, the queried object is matched with a tracking target in the gallery. The identification of a tracking target is an active research topic in the field of computer vision and has become popular with the prevalence of intelligent surveillance systems [1]. The ReID model
In this study, we propose Tiny Asymmetric Feature Normalized Network (TAFN-Net), a network that can learn generalized features using tiny parameters. The proposed method is a network that can subdivide and learn features at various scales. It is designed to adequately learn the discrimination power of both local and global features. TAFN-Net recalibrates global characteristics and reallocates importance per channel. Moreover, by designing an asymmetric regularization layer for regional characteristics, this model reduces the bias in the content statistics domain. As the number of regularization layers that are included increases, learning becomes more difficult; however, overfitting can be avoided.

When compared to previous studies, the primary contributions of this study are as follows: 1) A network based on domain generalization is designed to enable real-world applications. A generalized network composed of multiple branches is also designed. Local feature generalization is achieved using an asymmetric generalization layer. 2) Global feature generalization is achieved based on the use of the use of the instance-batch normalization and leaky ReLU (IBLR) squeeze-and-excitation (SE) module. Moreover, the importance of each channel of the global feature is recalibrated. 3) Thus, the dataset of the ReID system is effectively augmented and a practical person ReID application is demonstrated based on a new benchmark dataset. The rest of this research paper is structured as follows. Section 2 introduces the challenges and papers related to person ReID. Section 3 elaborately describes our proposed ReID model network structure. Section 4 introduces two large-scale benchmark dataset evaluation sets, a new benchmark dataset that was developed in this study, as well as an evaluation method. It also presents the implementation details. Conclusions and future work are described in Section 5.

II. RELATED WORKS
A. PERSON RE-IDENTIFICATION (ReID)

Convolutional neural networks (CNNs) have achieved superior performance improvement in all tasks across all subfields of computer vision, such as image classification and object detection. Accordingly, supervised image retrieval models have been proposed [9], [12]. Many of the latest ReID models are designed based on a residual network (ResNet) [13] as the backbone network. However, as these class-wise ReID models do not perform data preprocessing in the learning process, unnecessary noise is included. Studies that divide the human body into parts and match objects with identical IDs have been proposed [9], [14], [15]. In this method, the image is divided into several parts along the vertical direction and the segmented images are then input into the network. Rather than using global information, it is possible to improve the performance by effectively learning how to discriminate local features. However, because the division is performed based on absolute criteria, without considering the alignment of body parts, errors in recognition may exist.
even for the same person object. The limitations of the part-based ReID models were studied to solve the human pose variants [8], [16]. Openpose [17] extracts keypoints and estimates human poses. In the ReID model, the main key texture is selected based on the affine transformation of these keypoints. Preprocessing is performed selectively so that the ReID model can learn the texture. Experimentally, it was determined that combining local and global features yields optimal retrieval performance.

Subsequently, studies that actively considered the misalignment problem of part-based ReID models were proposed [18]. These new studies relied on a pictorial structure and matched the IDs using normalized poses to identify body parts. However, when the size of the test image in ReID is small, the process does not work well because it is difficult to estimate the pose. Therefore, when the ReID backbone network has sufficient generalization performance, there is no need for preprocessing steps, such as pose variation and alignment. Generally, well-generalized models can satisfactorily describe the appearance of global features. A representative model is SE network (SENet) [19], which was recently designed to be primarily used as a ReID backbone network. This network was originally designed for classification tasks; however, it can also be used for image retrieval methods based on metric learning [20]. If a ReID model is used as a backbone network using a lightweight deep neural network [21], [22], then it can be deployed on an edge device. However, this model yields relatively low performance when compared to the ReID-specialized model. Omni-scale network (OSNet) [23] has the advantage of being able to learn adaptively by diversifying omni-scaled features. Given that the number of parameters is small when compared to other backbone networks, this model can be easily used in practical applications. In TinyNet [24], the network-learned feature diversity is based on various preprocessing inputs, and similar to OSNet, a tiny network-based ReID model was developed.

B. NORMALIZATION FOR DEEP NEURAL NETWORK

Batch normalization (BN) is a method used for normalizing the feature map of a minibatch according to the batch rather than the global dimensions [25]. It is a simple yet powerful CNN learning method and has become the primary technique among training approaches. Reducing the batch size will nullify the effectiveness of the relatively accurate statistical estimation that is observed in large batch sizes. To compensate for this shortcoming, an instance normalization (IN) method [26] performs a BN-like calculation on a single sample. That is, the average and standard deviation are obtained, and normalization is performed on each channel of each image. A study conducted to improve the generalization ability of the model by integrating BN and IN was then proposed [27], [28]. Layer normalization (LN) [29] is a feature-level normalization method. In the case of BN, calculations are conducted on the entire batch, and the calculation is identical in each batch. However, in the case of LN, each characteristic is calculated independently. LN is independent of the batch size and has a strong effect on sequential data.

Group normalization (GN) [30] combines the advantages of both IN and LN. In other words, the mean and standard deviation are calculated by grouping channels. Despite its poor performance in image recognition, it is known to produce improved effects in relatively small batch sizes. The most recently proposed local context normalization [31] was designed as a normalization layer in which all feature maps were normalized based on the surrounding window and the filter of the corresponding group.

The introduced normalization layers can be used to optimally normalize the feature representation in person ReID. For example, a method yielding the optimal ReID performance was proposed as a method for reducing the variation in style statistics based on the use of IN of the initial layer [32]. This method is simple to implement and performs training and testing within a significantly short period.

III. PROPOSED METHOD

In this section, we describe the structure of our proposed TAFN-Net. The core regularization module of the proposed method is shown in Fig. 4. The complete system pipeline consists of a data augmentation part, a deep model (i.e., backbone), and a classifier. The BN-IN-GN (BIG) block is composed of modules within the deep model. The BIG block is a method used for normalizing the overall feature, and the IBLR SE block is used for recalibrating the global feature channel unit.
Algorithm 1 Algorithm for Tiny Asymmetric Feature Normalized Network

Input: Market1501 [2], DukeMTMC-reID [33], and GPP-reID dataset

Output: Weight $W'$ of the cross-domain image retrieval network

Training phase:
1: Initialize the weight $W$ of TAFN-Net $\phi$
2: Training from Input dataset (M, D)
   $L_{\text{total}} = \lambda_{\text{xent}} L_{\text{xent}} + (1 - \lambda_{\text{xent}}) L_{\text{tri}}$
3: Repeat process 2
4: if Weight $W$ is not updated then
5:   End of training phase
6: end if
7: return $W$

Fine-tuning phase:
8: Initialize TAFN-Net $\phi$’s weight according to $W$
9: Freeze weight of TAFN-Net except classifier
10: Fine-tuning the weight $W$ of TAFN-Net $\phi$ (GPP-reID)
   $L_{\text{total}} = \lambda_{\text{xent}} L_{\text{xent}} + (1 - \lambda_{\text{xent}}) L_{\text{tri}}$
11: Repeat process 10
12: if Weight $W$ is not updated then
13:   End of fine-tuning phase
14: end if
15: return $W$

A. NETWORK ARCHITECTURE FOR GENERALIZED FEATURE LEARNING

1) BIG BLOCK
In CNN, BN operation is a useful tool that can perform training with high stability. However, the generalization performance of the network cannot be guaranteed with the BN operation because the texture information of individual instances is consistently normalized.

Each image object in the ReID dataset has a style and the BN is normalized by addition in terms of overall minibatch. BN is based on the feature statistics of a batch sample that adds and normalizes all feature channels. Therefore, a lot of unique information about the image instances is lost compared to methods that normalize the samples in a single batch style [27], [34], [35]. IN-BN network [28] uses BN in combination with IN to obtain certain unique individual texture information to improve the performance. If BN is in charge of the overall normalization, IN can learn diverse features from one ID by using the ReID model. We focused on this point and designed the BIG block as shown in Fig. 2.

ReID performance is affected not only by the normalization method, but also by the batch size. When using IN, there is a conflict with BN because it calculates only one image of a minibatch and uses an individual distribution.

Using IN, all channels are induced to have different statistics, and through GN, certain features are bundled and normalized to ensure feature diversity.

When compared to the IN-BN (IB) block [28], the GN-BN (GB) block has a relatively mild structure with channel direction feature diversity. GN can be obtained as follows.

$$S_i = \{k | k_N = i_N, \left[\frac{k_C}{C/G}\right] = \left[\frac{i_C}{C/G}\right]\}$$  \hspace{1cm} (1)

The indices $i_C$ and $k_C$ are features in channels within the same group. $G$ is a hyperparameter of GN, which becomes LN when $G = 1$ and IN when $G = C$. GN is not superior to BN in relatively large batches in image recognition tasks; however, when used appropriately, it can contribute to the generalization of the ReID. If only IN is used in the normalization layer in network design, style variation is removed by directly normalizing the static features of the content image. Therefore, given that a certain degree of style normalization can be eliminated to maintain the discriminant of the ReID task, the BIG block was designed by appropriately including the GN operation in the IB block that overcomes these shortcomings.

The BIG block was applied to each layer in a multibranch-type network [23] that can effectively learn local features.

To match the number of channels, the output of the $1 \times 1$ convolution operation was used as the input to the BIG block. To achieve normalization to obtain fine features in the ReID person object, the BIG block was used in the feature extractor branch and configured in an asymmetrical form, as shown in Fig. 4. The key to improving performance in ReID work is to capture the characteristics of a person well in a low-resolution image. As a general approach, it is possible to search for similar people through the approximate appearance of the character’s characteristics (cloth color and texture information). However, it is difficult to search for hard cases based on texture information alone. (Objects with different IDs but similar texture) That is, ReID model should rely on fine-grained feature cues. Therefore, the BIG block is a normalization block that considers both the general approach and the difficult case of the ReID operation.

2) NONLOCAL ATTENTION
When considering the receptive field of CNN, even if a large kernel size is used, the area that the filter can observe at once is limited. Therefore, an iterative convolution operation is required to observe the appearance of the global features. Moreover, repetition of the same operation is inefficient and results in difficulty in optimization and the occurrence of a multihop dependency problem [36].

TAFN-Net was designed to operate a nonlocal operator inside the multibranch network. Features were extracted from each branch of the network to which multiple convolution filters were applied, and softmax outputs were calculated. The nonlocal operator of TAFN-Net was designed in a structure in which the intermediate softmax outputs were combined (element-wise multiplication) in the feature part, whereby the number of convolution filters was relatively sparse and was effective for learning global feature
diversity.

\[ f_i = x'_i \oplus \left( \text{softmax} \left( x_i^d \right) \otimes x_i^3 \right) \oplus \left( \text{softmax} \left( x_i^h \right) \otimes x_i^1 \right) \]

\[ + \left( \text{softmax} \left( z_i \right) \otimes x_i^1 \right) \otimes x_i^d + z_i \]  

(2)

In the above equation, \( x'_i \) is a skip connection identity feature, \( x_i \) indicates features from each of the diverse branch outputs, and \( z_i \) denotes features that were generated by the IBLR SE block. In the proposed method, nonlocal attention was calculated by \( 1 \times 1 \) convolution operations, and all the features were added element-wise, including the identity feature \( x'_i \), as shown in Fig. 4.

When only the neighborhood is calculated, receptive field diversity with a relatively small number of parameters can be achieved; the nonlocal operator of TAFN-Net can structurally compensate for this limitation.

B. IBLR SQUEEZE-AND-EXCITATION BLOCK FOR GLOBAL FEATURE DESCRIPTOR

SENet [19] can effectively learn the global descriptor. The appearance of a person undergoes a re-adjustment process according to the importance between the channels. The SE block can be easily applied to networks with a skip connection. In the proposed method, given that there is a skip connection of the TAFN module itself, the IBLR SE block was designed by modifying the SE module to each unit structure. The proposed method was inspired by the Network-in-Network architecture [37] and was designed to maximize the effect of feature recalibration and normalization by adding an IB block inside the module.

\[ f_{1+1\text{sc}} = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) \]  

(3)

First, the feature with a size of \( H \times W \times C \) is changed to \( 1 \times 1 \times C \). Here, \( u_c \) indicates the input feature of the \( c \) channel.

\[ f_{out} = \sigma \left( FC_2 \text{NormIB LR} \left( FC_1 \cdot f_{1+1\text{sc}} \right) \right) \]  

(4)

According to equation (4), features with a size of \( 1 \times 1 \times C \) were activated using leaky rectified linear unit (ReLU) \( \delta_{LR} \) and normalized using an IB block. The sigmoidal function \( \sigma \) is applied to the final output.

\[ x' = F_{\text{recalib}} \left( f_{out}, u_c \right) \]  

(5)

The scaled features were obtained by multiplying the \( H \times W \times C \) feature map \( u_c \) (before the squeeze) and \( f_{out} \). This was performed because all scale values have values that range between 0 and 1 after the excitation operation, but are scaled according to the importance of each channel. The number of middle nodes can be reduced through the reduction ratio \( r \) in the IBLR SE module. If the number of input feature maps is \( C \), the middle node can have a feature map equal to \( C/r \) to save the number of parameters. The IBLR SE block is positioned as shown in Fig. 4, which is independent from the convolution branch to avoid damaging the spatial information of each branch. The SE block calibrates global features and performs weight generalization simultaneously. To prevent knock out in the simple perceptron structure of the SE block, the leaky ReLu was used.

C. DATA AUGMENTATION: RANDOM ERASING, SEMANTIC COMBINATION

Data augmentation is essential as a learning method to avoid overfitting deep neural networks. Fig. 5 shows the data augmentation method that is applied in this study. In the proposed method, data is augmented in two ways.

The random erasing data augmentation method [38] is used to generate random noise or erase a region of interest after generating a window that has a random size and position. Given that ReID operates with the output bounding box of detection and tracking, background noise has a negative effect on the performance of the ReID retrieval model.

The proposed method considers the background complexity problem and adds augmented images in which background noise is removed to improve the foreground character features. We used human part segmentation [39] to segment the body into six parts (foot, hand, head, leg, shoulder, torso) and augmented the data with the combination that yielded the best ReID performance.

D. LOSS FUNCTION

1) IDENTIFICATION LOSS

The cross-entropy loss, which is commonly used in computer vision classification tasks, was considered as the identification loss in ReID. Using the softmax function, we estimated the probability as follows.

\[ \hat{p} (x_i) = \frac{\exp (x_i)}{\sum_{i=1}^{n} \exp (x_i)} \]  

(6)

\( \hat{p} (x_i) \) is the probability predicted by the softmax function and \( n \) is the number of samples. We adopted softmax cross-entropy loss \( L_{\text{sent}} \) as the identification loss, as shown subsequently.

\[ L_{\text{sent}} = - \sum_{i=1}^{n} p (x_i) \log \hat{p} (x_i) \]  

(7)

where \( \hat{p} (x_i) \) indicates the output predicted by the softmax function, \( p (x_i) \) indicates the target probability, and \( n \) is the number of samples.

2) TRIPLET LOSS

A triplet loss is a function of three variables. The triplet loss is primarily used for metric learning for clustering with high-dimensional results. It extracts features from objects of the same category and pushes features from other categories.

\[ L_{\text{tri}} = \sum_{i=1}^{n} \max \left( 0, D (a_i, p_i) - D (a_i, n_i) + m \right) \]  

(8)

where \( a_i, p_i, \) and \( n_i \) represent the descriptor features of an anchor, a positive, and a negative sample. \( D (a_i, p_i) \) indicates the distance between the anchor and the positive sample,
$D(a_i, n_i)$ indicates the distance between the anchor and the negative sample. $m$ is an element for including a margin. We set $m = 0.3$ for all the evaluated datasets.

3) TOTAL LOSS
By using a combination of identification and triplet losses in the ReID task, the advantages of both losses can be obtained simultaneously. In this study, we used a combination of the two loss functions as follows.

$$L_{\text{total}} = \lambda_{\text{xent}} L_{\text{xent}} + (1 - \lambda_{\text{xent}}) L_{\text{tri}},$$

(9)

where $\lambda_{\text{xent}}$ is the balancing weight of the two losses. In our experiments, $\lambda_{\text{xent}} = 0.8$ yielded the best performance.

E. OVERALL FRAMEWORK
ReID task focuses on identifying people across different scenes in video surveillance, which is commonly formulated as a ranking task via the feature-based ReID model approaches.

In general, the ReID surveillance system goes through five procedures, as follows. 1) Surveillance system collects raw images. 2) Create a bounding box by using detection and tracking algorithms. 3) Annotate raw images. In the closed world (i.e., for a specific dataset) environment, classification or verification ReID is used, and in the case of the open world, unsupervised ReID is used. 4) Train the ReID model.
FIGURE 7. Structure of the surveillance system based on the proposed person ReID model (TAFN-Net).

through a refined dataset. 5) Retrieved ranking that compares query-to-gallery similarities and lists them in descending order. A re-ranking algorithm (ranking optimization) can be considered as post-processing, but it is not suitable for actual video surveillance environments.

The complete framework of the proposed model is shown in Fig. 7. First, to construct the Gwangju Institute of Science and Technology (GIST) Practical Person ReID (GPP-reID) dataset, raw data images were collected from six surveillance cameras installed at GIST University, as shown in Fig. 6. We performed object detection using Mask regional CNN [41]. To achieve an application system that is capable of open-world person ReID, we adopted multi-object tracking. Gaussian mixing probability hypothesis density (GMPHD)-occlusion group management [40] is effective against object occlusion and false negative detection in the detection environment as it uses the GMPHD filter, which is based on hierarchically structured data association. In the GPP-reID dataset, images from different cameras and similar labeled data were bound together and designated as a gallery. The data format is ID/CamScene/Frame, and it is configured similarly to the data format of Market1501 [2]. For the first system operation, TAFN-Net was trained using a public dataset and the classifier was fine-tuned using a single-scenario dataset of GPP-reID. When $K > 1$, the adapted boost was used for learning based on the random sampling of datasets from $K - 1$ to $K$, as shown in Fig. 7. At the final scenario $K$, the weight of TAFN-Net was updated with an exponential moving average, as shown by the following equation.

$$x_k \leftarrow ax_k + (1 - a)x_{k-1}$$

Here, $x_k$ is the weight updated by the dataset up to scenario $K$, and $a$ is the weight that determines the importance of the data of the current and previous scenarios. Surveillance ReID was performed by the boosted ensemble TAFN-Net model according to scenario $K$.

IV. EXPERIMENTS

We present the results of a quantitative evaluation performed on the proposed TAFN-Net model in this section. First, we used a benchmark public dataset; then, the new benchmark GPP-reID dataset was introduced. Subsequently, Sections B and C describe the ReID evaluation metrics and implementation details, respectively. Section D provides a comparison with other methods. The datasets used in the
experiment were the extensively used, large-scale person ReID dataset and the GPP-reID dataset, which can evaluate the performance of domain generalization. The datasets include the natural deformable poses of a person object by changing the intensity of lighting and perspective in the surveillance environment. All datasets consisted of a training set and test set (query and gallery) with a total of 12,936 and 23,100 image datasets with 751 and 750 IDs, respectively.

A. DATASET FOR PERSON ReID

1) Market1501

The Market1501 [2] dataset consists of 1,501 IDs. All datasets consist of 32,668 images, excluding the query set, with labeled bounding boxes. Data acquisition was conducted by setting a nonoverlapping cross camera based on the six surveillance cameras that were installed at Tsinghua University. The datasets consisted of a training set and test set (query and gallery) with a total of 12,936 and 23,100 image datasets with 751 and 750 IDs, respectively.

2) DukeMTMC-reID

The Duke Multitracking Multicamera ReID (DukeMTMC-reID) dataset is a dataset that was created for multi-object tracking. It is a sub-dataset of DukeMTMC [62], which was created for the identification of people from images. The data acquisition method was implemented with a nonoverlapping cross-camera setting based on eight surveillance cameras that were installed at Duke University. All datasets were separated similarly to the Market1501 dataset. This dataset consists of 16,522 training sets and 2,228 and 17,661 test sets (query and gallery), respectively, comprising 1,404 IDs. The DukeMTMC-reID dataset includes pedestrian images with different IDs and contains multiple similar images that are difficult to recognize because of occlusion.

3) GPP-reID

The GPP-reID dataset consists of images of people that were acquired using six surveillance cameras at GIST University. Using this dataset, we can test a total of eight sets of image retrieval in four scenarios with two trials. There are 127 queries with a total of eight IDs, and the gallery set consisted of a total of 13,064 image datasets. The GPP-reID dataset was created to normalize and fine-tune a model trained with a large public ReID dataset. It is possible to estimate the performance of domain generalization, and an example is shown wherein a person image retrieval application was developed in a place equipped with a specific surveillance environment.
TABLE 2. Performance comparison of convolutional neural network (CNN) models trained from scratch. We conducted experiments without fine-tuning and data augmentation.

| Networks   | Market1501 | DukeMTMC-reID | Params |
|------------|------------|---------------|--------|
|            | R@1 | mAP | R@1 | mAP |          |
| SqueezeNet | 65.4 | 51.7 | 58.2 | 39.6 | 0.74M    |
| ResNet18   | 74.3 | 65   | 60.9 | 42.3 | 11.2M    |
| ShuffleNetv2 | 79.8 | 56.8 | 67.8 | 48.1 | 1.25M    |
| DenseNet121 | 83.1 | 63.5 | 73.2 | 55.1 | 6.95M    |
| MobileNetv2 | 83.2 | 60.1 | 71.3 | 50.3 | 2.22M    |
| MLFN       | 83.4 | 66   | 69.6 | 48.6 | 32.5M    |
| ResNet50   | 84.9 | 67.2 | 74.8 | 55.4 | 23.5     |
| ResNet101  | 85.6 | 72   | 76.4 | 59   | 42.5M    |
| Inceptionv4 | 88.2 | 79.5 | 77.4 | 60.7 | 41.1M    |
| OSNet      | 92.8 | 81.3 | 83.7 | 67.9 | 2.49M    |
| TAFN(ours) | 93.6 | 83.2 | 85   | 67.1 |          |

TABLE 3. Performance according to the type of CNN model in cross-validation setting. M denotes Market1501 and D denotes DukeMTMC-reID datasets.

| Networks   | M → D | mAP | D → M | mAP |
|------------|-------|-----|-------|-----|
| DenseNet121 | 11.7  | 5   | 32.7  | 11.6 |
| Inceptionv4 | 12.4  | 6.9 | 36    | 11.2 |
| ResNet50   | 14.6  | 5.2 | 34.7  | 13.5 |
| ResNet101  | 14.9  | 8.1 | 35.6  | 13.3 |
| SqueezeNet | 15.3  | 7.3 | 30.1  | 10.8 |
| MLFN       | 15.8  | 8.5 | 34.3  | 12.7 |
| MobileNetv2 | 17.1  | 7.5 | 35.1  | 14.2 |
| ShuffleNetv2 | 17.2  | 7.2 | 36.8  | 13.1 |
| OSNet      | 24.1  | 8.2 | 36.1  | 14   |
| TAFN(ours) | 25.3  | 12.8| 36.4  | 13.9 |

The GPP-reID dataset has the advantage of being composed of datasets for each scenario. Each scenario simulates an item-theft situation, and the camera topology can be estimated by identifying the trajectory of the human object. GPP-reID consists of images of various scales with the same ID (Fig. 9); therefore, the scalability performance of the ReID model can be assessed. It also contains several occlusions that can occur naturally because of objects such as trees and cars.

**B. EVALUATION METRICS**

Metric learning approaches are implemented by embedding two feature vectors with the same ID. In general, distance measurement methods used in the ReID model include the measurement of Euclidean, Hamming, and Manhattan distances. Cosine distance measurement was primarily used in person image retrieval tasks [63]. The cosine distance metric learns the optimized feature space by simply parametrizing the classification system of the softmax output. In this study, the distance between the feature vectors A and B calculated from the two input images was calculated as follows.

\[
A \cdot B = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{11}
\]

1) **CUMULATIVE MATCHING CHARACTERISTIC (CMC)**

Curve evaluates the ReID model performance. For each query, the ReID model can sort the gallery set according to the cosine similarity and measure the CMC top-k accuracy.

\[
Acc_k = \begin{cases} 
1 & \text{if top-}k \text{ranked samples same ID with } q \\
0 & \text{otherwise} 
\end{cases} \tag{12}
\]

CMC computes the closest image in the gallery set corresponding to a query image.

2) **MEAN AVERAGE PRECISION (mAP)**

Is a criterion used for evaluating the ReID model performance. It is determined by calculating the average of the average precision values.

\[
mAP = \frac{\sum_{q=1}^{Q} \text{AveP}(q)}{Q} \tag{13}
\]

If a query image is not matched in the gallery set, the precision value becomes zero. In this study, to evaluate the actual performance of the proposed TAFN-Net model and compare the results with other ReID models, the accuracy was calculated and evaluated using mAP and Rank-1 accuracy.

**C. IMPLEMENTATION DETAIL**

In this study, we implemented the proposed method using PyTorch, a deep learning framework based on Python 3.6, and the graphical user interface (GUI) was constructed using PyQt4.

First, to overcome the occlusion problem of the TAFN-Net model and to increase generalization [38], data augmentation was performed with random erasing and semantic segmentation of body parts. For TAFN-Net training, Adam optimizer [64] was employed to train the TAFN-Net model for up to 120 epochs, and the learning rate was adjusted according to a predefined schedule. The initial learning rate was set to 0.00003, and the learning rate was multiplied by 0.1. In the test phase, we obtained a 512-dimensional pedestrian descriptor as the output of TAFN-Net. Rank was calculated by comparing the 512-dimensional feature distance between one selected query image and all the images in the gallery set, and by arranging them in the order of increasing distance from the query image.

**D. COMPARISON WITH OTHER METHODS**

We evaluated the proposed TAFN-Net model on two large-scale benchmark datasets and the new dataset, GPP-reID, and compared the results with those of existing methods. For the comparison, we calculated the number of parameters and performances of the backbone networks that used the global feature learning-based ReID models. This study aimed to implement a practical surveillance application of the ReID model. Therefore, reranking [42], which is a
TABLE 4. Experimental results of Tiny Asymmetric Feature Normalized Network (TAFN-Net) method in single domain.

| Methods                     | Market1501 R@1 mAP | DukeMTMC-reID R@1 mAP |
|-----------------------------|---------------------|------------------------|
| $B$ (Baseline)              | 92.8                | 81.3                   |
| $B + IBLR$                  | 94.2                | 83.1                   |
| $B + IBLR + BIGnorm$        | 93.6                | 82.9                   |
| $B + IBLR + BIG + NL$       | 93.6                | 83.2                   |
| $B + IBLR + BIG + NL + Aug$ | 94.7                | 86.2                   |

postprocessing technique in the ReID model that has a long inference time, was not considered, and all evaluations were composed of a single query.

We compared the proposed TAFN-Net model with global feature learning-based models on the Market1501 dataset and the results are listed in Table 1. It can be observed that the proposed model exhibits superior performance with respect to the number of parameters. mAP and Rank-1 accuracy values of 86.2 and 94.7, respectively, were obtained. When the same evaluation method was applied to the DukeMTMC-reID dataset, the mAP and Rank-1 accuracy values were 74.8 and 85.9, respectively.

The performance of person ReID was affected considerably by the network normalization ability and diverse feature learning rather than the network types and the number of parameters, which can be confirmed from the experimental results. However, the simple image classification task exhibits a large variation between classes in the dataset. Moreover, it can be confirmed that the person ReID task relies on fine-grained cues or texture features within the same person object. Table 2 lists the results of initializing the various networks proposed for the image classification task without ImageNet fine-tuning and based on learning with ReID from the beginning.

Given that TAFN-Net designed and configured various normalization layers inside the network, supervised cross-data evaluation was performed to measure the normalization ability of TAFN-Net. Unsupervised methods are unsuitable for practical ReID models. Thus, we compared them with various deep learning networks with supervised settings and the results are listed in Table 3.

TAFN-Net is suitable for practical ReID surveillance applications because the model size was relatively small; moreover, superior general performance and cross-validation ability were demonstrated.

E. ANALYSIS OF TAFN-NET MODEL

For comparison, experiments were conducted by adding the element methods of TAFN-Net based on OSNet [23], which is a multibranch-type network. For ReID evaluation, general and cross-validation evaluations were performed, and the results are listed in Tables 4 and 5, respectively. The baseline model was trained from the beginning on the public person ReID dataset without ImageNet fine-tuning. The IBLR SE block achieved improvements in mAP. Conversely, in the cross-validation setting, the nonlocal operator resulted in a reduction in mAP at specific settings, as listed in Table 5. Experimental results show that the performance of the ReID model can be improved maximally when the TAFN-Net method is performed together with data augmentation, which can be noted based on the results listed in Tables 4 and 5.

The proposed TAFN-Net model was trained by combining two loss functions. In our work, the weight $\lambda_{xent}$ was set according to equation (9) to balance cross-entropy and triplet losses. Table 6 lists the effect on model performance according to the weight $\lambda_{xent}$ value. As a result of the experiment, the optimal performance of TAFN-Net was obtained by setting the weight $\lambda_{xent} = 0.8$.

TABLE 5. Experimental results of the TAFN-Net method in the cross-domain. M denotes Market1501, D denotes DukeMTMC-reID, and G denotes Gwangju Institute of Science and Technology Practical Person re-identification (ReID) datasets.

| Methods                     | Market1501 R@1 mAP | DukeMTMC-reID R@1 mAP |
|-----------------------------|---------------------|------------------------|
| $B$ (Baseline)              | 58.2                | 41.4                   |
| $B + IBLR$                  | 59.7                | 40.1                   |
| $B + IBLR + BIGnorm$        | 60.8                | 44.2                   |
| $B + IBLR + BIG + NL$       | 60.7                | 41.9                   |
| $B + IBLR + BIG + NL + Aug$ | 62.9                | 43.1                   |

$2$-scene fine-tuning
$x_2 \leftarrow a x_2 + (1 - a) x_1$

TABLE 6. ReID performance results according to triplet and cross-entropy ratios.

| $\lambda_{xent}$ | Market1501 R@1 mAP | DukeMTMC-reID R@1 mAP |
|------------------|---------------------|------------------------|
| $\lambda_{xent} = 1.0$ | 94.3                | 84.6                   |
| $\lambda_{xent} = 0.8$ | 94.7                | 86.2                   |
| $\lambda_{xent} = 0.6$ | 94.5                | 85.5                   |
| $\lambda_{xent} = 0.5$ | 93.7                | 85.2                   |
| $\lambda_{xent} = 0.2$ | 93.9                | 85.1                   |
| $\lambda_{xent} = 0.1$ | 94.9                | 84.8                   |

TABLE 7. Experimental results of the degree to which human body parts affect ReID feature descriptors.

| Body parts segmentation | Market1501 R@1 mAP |
|-------------------------|---------------------|
| Head                    | 2.8                 |
| Torso                   | 21.3                |
| Shoulder                | 4.1                 |
| Leg                     | 30.6                |
| Hand                    | 2.9                 |
| Foot                    | 3.2                 |
| Torso+Leg               | 48.6                |
| Head+Torso+Leg+Foot     | 67.3                |
| Whole body              | 73.3                |

In the cross-validation setting, the nonlocal operator resulted in a reduction in mAP at specific settings, as listed in Table 5. Experimental results show that the performance of the ReID model can be improved maximally when the TAFN-Net method is performed together with data augmentation, which can be noted based on the results listed in Tables 4 and 5.
F. ABLENTION STUDY

We evaluated the effect of partial features on the ReID task. The proposed method is a study that explores an effective method in the supervised cross-validation setting by maximally extracting feature diversity from the public dataset. Therefore, to analyze the characteristics of a person object, the effects of all the body parts (Fig. 5) on the ReID model were evaluated; these results are listed in Table 7.

Segmentation of human parts [39], a new dataset in which the background has been removed, was created by segmenting the region into seven parts, namely, the head, torso, shoulder, leg, hand, and foot. As a result of the experiment, Rank-1 accuracy and mAP values of the head part were 2.8 and 0.8, respectively. Thus, it can be confirmed that the features helpful for the ReID task are sparse.

Similarly, results showed that the hand and foot did not contain any remarkable features. It can be confirmed that the main key factors for improved ReID performance exist in considering a combination of torso and leg features.

In this study, a new dataset where the background was removed was created and used for model training. One-half of the total augmentation images were generated using a combination of the entire body with the head + torso + leg + foot features, and the remaining parts were generated by the random erasing method [38].

V. CONCLUSION

We proposed a specialized network for feature normalization called TAFN-Net and a framework to perform person image retrieval by applying it to a practical surveillance environment. TAFN-Net could effectively extract the specified global feature information by including the designed nonlocal operator and IBLR SE block. Experimental results showed that the proposed network could satisfactorily learn the discriminative multiscale and generalization features. Data augmentation was performed to improve the generalization performance by analyzing the degree to which each body part of a person affects the ReID task. Superior performance was achieved using a relatively small model on the benchmark dataset for ReID. Cross-validation performance was assessed on GPP-reID, which is a new benchmark dataset that includes different scenarios. If the time of appearance of objects in the dataset is annotated, the camera topology can be estimated. Particularly, in terms of model performance, the model is expected to improve the generalization performance based on reweighting, as shown in Fig. 7. TAFN-Net can automate surveillance systems and display the results via a GUI interface. The results showed that several ReID research models can be applied to real-world monitoring systems.

REFERENCES

[1] L. Zheng, Y. Yang, and A. G. Hauptmann, “Person re-identification: Past, present and future,” 2016, arXiv:1610.02984.
[2] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable person re-identification: A benchmark,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1116–1124.
[3] J. Li, S. Zhang, Q. Tian, M. Wang, and W. Gao, “Pose-guided representation learning for person re-identification,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 2, pp. 622–635, Feb. 2022.
[4] W. Li, R. Zhao, T. Xiao, and X. Wang, “DeepReID: Deep filter pairing neural network for person re-identification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 152–159.
[5] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, “Person re-identification by local maximal occurrence representation and metric learning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 2197–2206.
[6] R. R. Varior, M. Haloi, and G. Wang, “Gated Siamese convolutional neural network architecture for human re-identification,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2016, pp. 791–808.
[7] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, “A Siamese long short-term memory architecture for human re-identification,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2016, pp. 135–153.
[8] L. Zheng, Y. Huang, H. Lu, and Y. Yang, “Pose-invariant embedding for deep person re-identification,” IEEE Trans. Image Process., vol. 28, no. 9, pp. 4500–4509, Apr. 2019.
[9] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, “Spindle net: Person re-identification with human body region guided feature decomposition and fusion,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1077–1085.
[10] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, “Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 1004–1013.
[11] J. Wang, X. Zhu, S. Gong, and W. Li, “Transferable joint attribute-identity deep learning for unsupervised person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 2275–2284.
[12] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, “Person re-identification by multi-channel parts-based CNN with improved triplet loss function,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 1335–1344.
[13] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.
[14] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, “Deep representation learning with part loss for person re-identification,” IEEE Trans. Image Process., vol. 28, no. 6, pp. 2860–2871, Jun. 2019.
[62] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi, “Performance measures and a data set for multi-target, multi-camera tracking,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 17–35.
[63] N. Wojke and A. Bewley, “Deep cosine metric learning for person re-identification,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Mar. 2018, pp. 748–756.
[64] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2017, arXiv:1412.6980.

HOYEON AHN received the B.S. degree in industrial engineering from Ajou University, Suwon, Republic of Korea, in 2016, and the M.S. degree in electrical engineering from Kyungpook National University, Daegu, Republic of Korea, in 2018. He is currently pursuing the Ph.D. degree with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea. His current research interests include artificial intelligence, machine learning, and pattern recognition.

YOOJIN HONG received the B.S. degree in computer science from Sookmyung Women’s University, Seoul, Republic of Korea, in 2016, and the M.S. degree from the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), Gwangju, Republic of Korea, in 2018, where she is currently pursuing the Ph.D. degree. Her current research interests include artificial intelligence, face recognition, face detection, and android.

HYUNGUK CHOI received the B.S. degree from the Department of Electronic Engineering, Dongguk University, Seoul, Republic of Korea, in 2012, and the Ph.D. degree from the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea, in 2020. He is currently working as the Head of Research Institute with Satrec Initiative Analytics, Daejeon, Republic of Korea. His research interests include computer vision, artificial intelligence, and machine learning.

JEONGHwan GWAK (Senior Member, IEEE) received the Ph.D. degree in machine learning and artificial intelligence from the Gwangju Institute of Science and Technology (GIST), Gwangju, South Korea, in 2014. From 2002 to 2007, he worked for several companies and research institutes as a researcher and a chief technician. From 2014 to 2016, he worked as a Postdoctoral Researcher at GIST. From 2016 to 2017, he worked as a Research Professor. From 2017 to 2019, he was a Research Professor at the Department of Radiology, Biomedical Research Institute, Seoul National University Hospital, Seoul, South Korea. In 2019, he joined the Korea National University of Transportation as an Assistant Professor, and he is currently the Director of the Algorithmic Machine Intelligence Laboratory. His current research interests include deep learning, computer vision, image and video processing, evolutionary computation, optimization, and relevant applications of medical and visual surveillance systems.

MOOGU JEON (Senior Member, IEEE) received the B.S. degree in architectural engineering from Korea University, Seoul, Republic of Korea, in 1988, and the M.S. and Ph.D. degrees in computer science and scientific computation from the University of Minnesota, Minneapolis, MN, USA, in 1999 and 2001, respectively. As a Postgraduate Researcher, he worked on optimal control problems at the University of California at Santa Barbara, Santa Barbara, CA, USA, from 2001 to 2003; and then moved to the National Research Council of Canada, where he worked on the sparse representation of high-dimensional data and image processing until July 2005. In 2005, he joined the Gwangju Institute of Science and Technology, Gwangju, Republic of Korea, where he is currently a Full Professor with the School of Electrical Engineering and Computer Science. His current research interests include machine learning, computer vision, and artificial intelligence.