Journal of Information Hiding and Multimedia Signal Processing ©2024 ISSN 2073-4212
Ubiquitous International Volume **15**, Number **3**, September **2024**

A Modification of the Temporal Group Attention Method on Super-Resolution Video for Vehicle Number Plate Detection

Budi Setiyono[1], Dwi Ratna Sulistyaningrum[2], Ario Fajar Pratama[3]
Ridho Nur Rohman Wijaya[4]

Departement of Mathematics
Institut Teknologi Sepuluh Nopember Surabaya, Indonesia
[1]masbudisetiyono@gmail.com,[2]dwratna@gmail.com,[3]ariofajarpratama@gmail.com
[4]ridhonurrohmanwijaya@gmail.com

Correspond Author : masbudisetiyono@gmail.com
Received June 2024; revised July 2024, accepted July 2024

Abstract. *Today, the development of smart cities is relatively rapid, and one of the components of a smart city is the Intelligent Transportation System (ITS). Vehicle detection and assistance on the road is part of ITS, which can be done by identifying the vehicle's license plate. In the following study, the authors conducted the detection and assistance of motorized vehicle license plates by modifying the TGA network. Altering the TGA network on super-resolution video increases the accuracy and speed of detecting vehicle license plates. Furthermore, with network modification, the temporal information retrieved can be more balanced and have a more dynamic structure. We made modifications in several parts : (i) using the PDC (Pyramid Deformable Convolution) method for registration, where this process did not exist in the previous TGA. (ii) on Temporal Grouping, group division is made dynamic depending on the number of input frames. (iii) on the Intra Group Fusion Module, using the MSTC (Mixed Spatial-Temporal Convolution) method, while on the previous TGA, using 3D convolution. (iv) reducing the number of repetitions of the method DSC (Dense Skip Connections) in feature extraction and reducing the use of GCSC (Group Convolution with Skip Connections) only once in the Intra Group Fusion Module process. According to the experiment, our modification can boost accuracy by up to 4.23% while maintaining the same computation time. Therefore, this research contributes positively to the problem of vehicle license plate detection.*
Keywords: : Smart City, Intelligent Transportation System, TGA Modification.

1. **Introduction.** Smart cities are rapidly growing, and one important aspect is the Intelligent Transportation System (ITS). ITS has several research areas that can be explored, such as improving video frame extraction [1], developing image classification methods [2], and building vehicle identification systems [3]. In particular, detecting and identifying vehicles on the highway is a crucial component. Vehicle identification can be accomplished by identifying the vehicle's license plate [3–5]. Vehicle number plates are a unique and visible identity and are a source of information regarding detailed specifications and vehicle ownership. The author uses a digital image processing approach to detect vehicle license plates. The image obtained must have a high resolution for identification to be easy. However, obtaining vehicle license plate data from CCTV recordings typically results in low-resolution images, making identification

challenging [6–8]. In digital image processing, the image quality will be significantly influenced by the level of resolution. The higher an image resolution, the greater its pixel density and the richer the visual details. Another way to increase image resolution is by using super-resolution techniques [9, 10].

Rapid scientific progress has made many super-resolution methods use a deep learning approach in their applications. Deep neural networks are a data-driven approach that examines the characteristics of input patterns. Continuous advances in deep learning research show superior speed and accuracy compared to traditional computer vision algorithms [11]. As a result, data scientists and practitioners often adopt this model in the machine learning industry. Therefore, extending this deep learning approach in super-resolution based image reconstruction applications is critical.

Most video super-resolution techniques generally depend on either time-based or space-based temporal information. In a time-based system, frames are treated as time series data and sent over the network [12, 13]. This approach frequently employs an iterative method, using the previously estimated high-resolution frame as input to aid in reconstructing the next low-resolution frame. However, since it cannot analyze multiple frames concurrently, this method can only utilize a limited amount of temporal information from earlier frames and operates at a slower pace.

Super-resolution video with a space-based approach to low-resolution target frames considers the large size of adjacent frames as an additional information resource [14]. Several techniques are used to combine this information, including selective fusion [15], deep fusion [16], or 3D convolutional neural networks [17]. Compared to time-based techniques, this supports maintaining correlation between frames while taking advantage of parallel computations. Furthermore, video sequences represent spatial-temporal information in three dimensions. Therefore, a model of merging spatial-temporal information in neural networks using 3D-CNN provides a natural and economical solution for super-resolution video [18]. The latest 3D-CNN network stacks 3D layer by layer for super-resolution video.

In license plate detection, super-resolution video techniques enhance the resolution of detected objects, transforming low-resolution frames into high-resolution ones. This improvement is expected to increase the accuracy of license plate localization. In our research, we use the YOLOv3 method to detect vehicle license plates because of its fast and effective performance in license plate detection as in the research [19]. That research also used YOLOv3 to detect similar problems for the license plate recognition. In addition, the method also has a fundamental advantage, which is a simple architecture which makes it rapid in detection and evidenced by the stable mAP with fast inference time.

Based on the description above, if the detection process produces a negative image or many license plates are not detected, it will affect the character recognition process, such as low character recognition accuracy. Therefore, our research focus on the performance of super-resolution video in detecting license plate locations. The video super-resolution method is Temporal Group Attention (TGA) with a temporal super-resolution technique to obtain higher image quality [20]. In addition to improving visual quality, TGA is also used to get more detailed information from each frame to detect the movement of plate objects and produce more accurate detections. Our research also uses feature extraction and attention module with the Mixed Spatial Temporal Convolution approach, which is proven to handle detail recovery better in [15].

2. **Briefly Overview of Related Work.** Research conducted by [6] with title Number plate recognition on vehicles using YOLO-Darknet. This research discusses the introduction of vehicle license plate characters into the text. The study results indicate

that the accuracy of credit using the YoloV3 method for numeric characters and letters of the alphabet increases to around 88% and 97.6%, respectively, when pre-processing is carried out. This study conducted experiments under various lighting conditions ranging from -75 to +75. YoloV3 demonstrated significant accuracy in performance.

Sun [15] demonstrates video super-resolution using mixed spatial-temporal convolution and selective fusion. This research proposes a video super-resolution network with Mixed Spatiotemporal Convolution, which enables 2D CNNs to extract deeper spatial features and capture temporal information between consecutive frames using 3D convolution as an additional component to enhance detail reconstruction. Experimental results show that MSTC significantly outperforms the performance of 2D or 3D CNNs, and the selective feature fusion strategy better handles detail recovery. Additionally, comparisons across various test videos with different types of motion, scenes, and input sizes demonstrate that this approach achieves acceptable performance when processing multiple SR videos. However, the performance could be better for small textures; some minor details still need to be more transparent and straightforward than the ground truth.

Another research completed by [20], Video super-resolution with Temporal Group Attention. In his research, Isobe proposes a new method for hierarchically combining temporal information in video super-resolution. This approach utilizes spatial and temporal information across all frames to enhance the details of low-resolution frames. However, this research still needs to explain in detail the analysis of motion compensation when the object in each frame experiences different movements.

Khan proposed a method called DSTnet: Deformable Spatio-Temporal Convolutional Residual Network for Video Super-Resolution [21]. This research proposes a method consisting of $(2 + 1)D$ spatiotemporal residual convolution blocks with deformable convolution layers to simultaneously utilize spatial and temporal information. Experimental results confirm that DSTnet can effectively catch and model complex motion between frames in the Vid4 benchmark dataset. The proposed method is evaluated using SSIM and PSNR, achieving SSIM of 0.795 and PSNR of 26.39 dB. Additionally, the proposed method has fewer parameters to learn during training, making it computationally leaner and exhibiting fast learning capability.

3. **Methodology.** This method takes a sequence of $N = 2n + 1$ LR frames $\{I_{t-n}, ..., I_t, ..., I_{t+n}\}$ as input. Where I_t is the target frame. The SR video aims to release a version of HR I_t. The framework consists of three main modules: motion compensation, deep spatial-temporal feature extraction, and selective feature fusion. Figure 1 shows the overall structure of the proposed network.

3.1. **Motion compensation.** Firstly, we extracted the first flat feature from each frame adjacent to $\{F^i_{t-n}, ..., F^i_t, ..., F^i_{t+n}\}$ by utilizing shared 3×3 convolutional layers $C_i(\cdot)$. Then we employ a Pyramid Deformable Convolution (PDC) as in [22] to align each neighboring frame to a single target frame F_t at the feature level, as illustrated in Figure 2. This involves applying pyramid processing to the shallow features of each neighboring frame and the target frame to generate multi-scale features. At the l-th level of the pyramid, offsets and aligned features are predicted using $\times2$ offset upsample and features aligned from above at the $(l-1)$-th level, we get:

$$\Delta P^l_{t-n} = C_f([F^i_{t-n}, F_t], (\Delta P^{l-1}_{t-n})^{\uparrow 2}) \tag{1}$$

$$Z_{t-n} = C_g\left(DConv(F^i_{t-n}, \Delta P^l_{t-n}), \left((F^i_{t-n})^{l-1}\right)^{\uparrow 2}\right) \tag{2}$$

where:

1. ΔP^l_{t-n} is the resulting offset for the deformable convolution at the l-level,

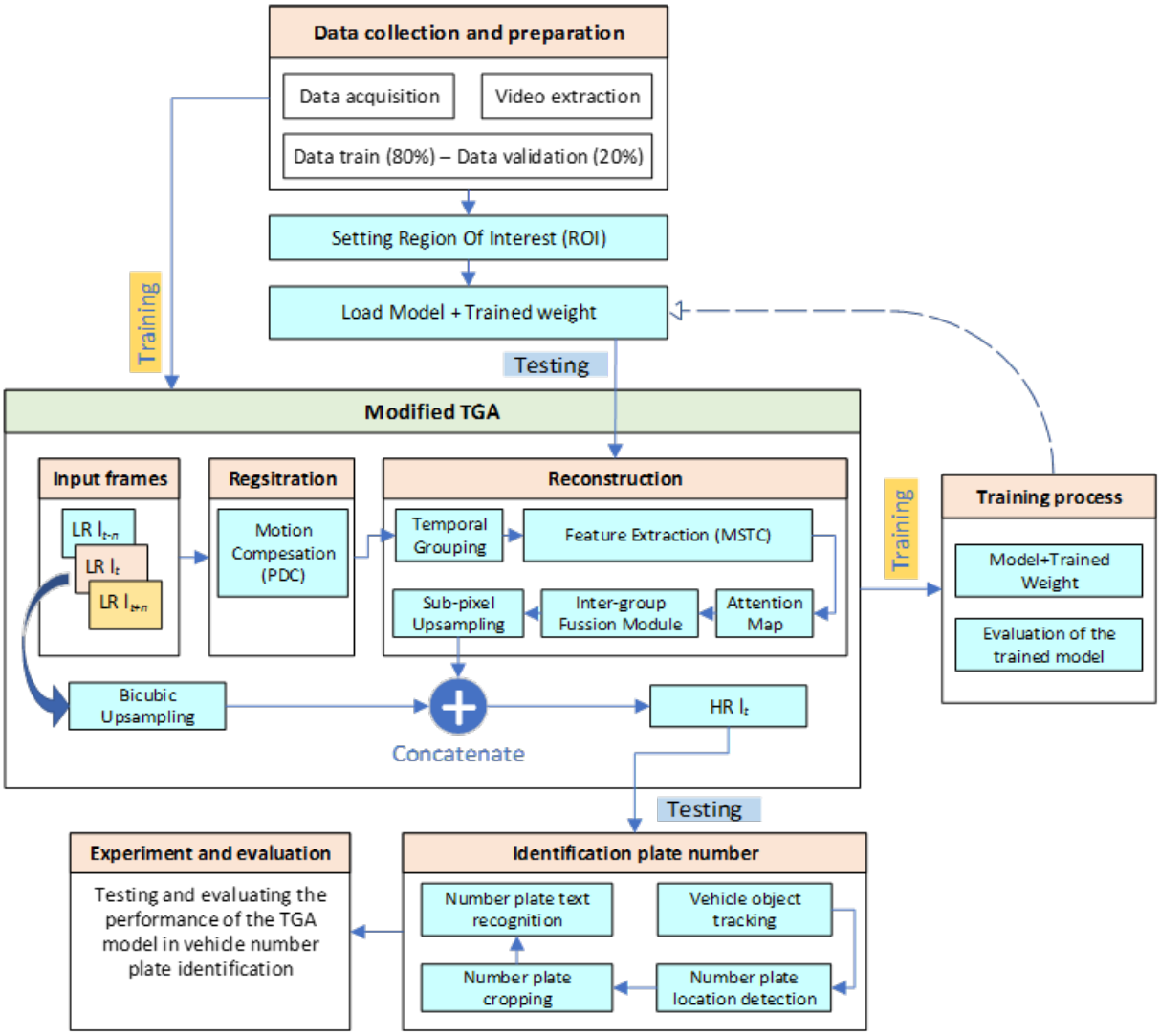

FIGURE 1. Research Work Block Diagram

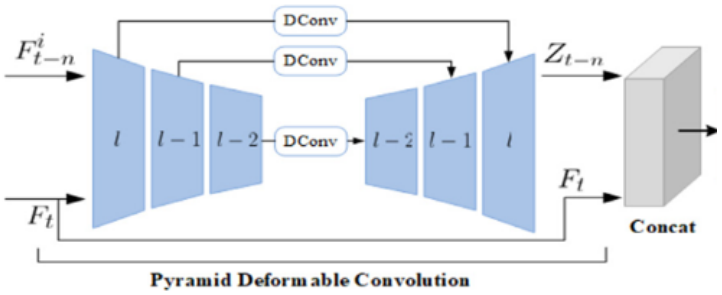

FIGURE 2. Motion Compensation Architecture

2. $[\cdot,\cdot]$ is a concatenation operation,
3. $(\cdot)^{\uparrow 2}$ is an upscale process with a factor of 2,
4. $(F^i_{t-n})^{l-1}$ is the generated features at the $(l-1)$-th level,
5. C_f and C_g are general function consisting 3×3 convolution layers,
6. $DConv$ means the deformable convolution

The visual informativeness of each frame and avoiding harmful features entering into the merging process, we used the attention module to calculate the correlation of each

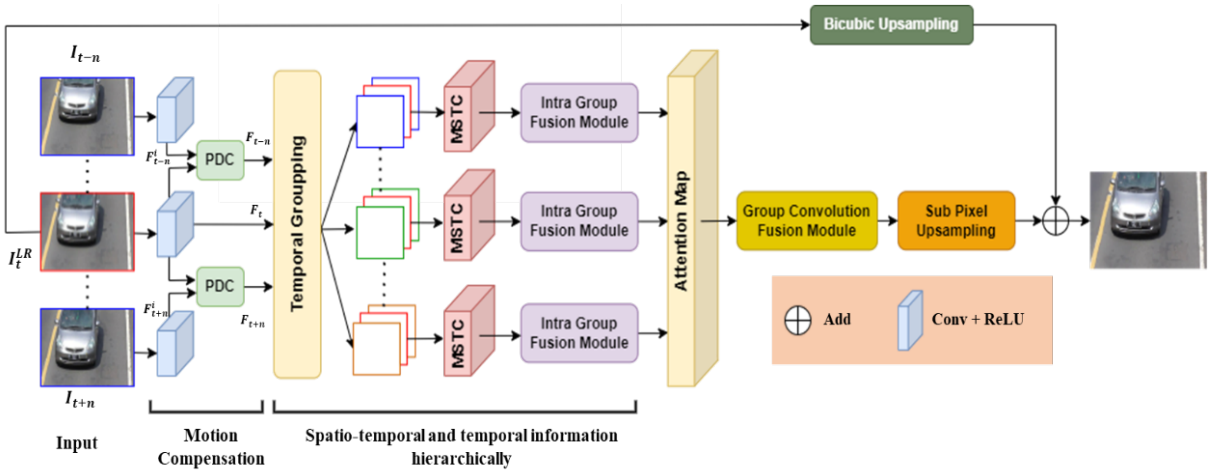

FIGURE 3. Entire network architecture with PDC and MSTC extensions to TGA modification.

element between the features at each location, which shows how informativeness to reconstruct target frame. The $\{Z_{t-n}, ..., Z_{t-1}, Z_{t+1}, ..., Z_{t+n}\}$ aligned features fed into the $C_a(\cdot)$ attention module. In particular, considering the aligned features of Z_{t-n}, the feature map of F^l_{t-n} is obtained through the attention module.

$$F^l_{t-n} = C_a([Z_{t-n}, F_t]) \tag{3}$$

Details of the concern module are shown in Figure 3. This module was first performed by combining the features of Z_{t-n} with F_t, then obtained the attention of 64-channel F^l_{t-n} maps.

3.2. **Temporal Grouping.** Two adjacent sets of N images were grouped based on the duration of the target frame, and the original sequence was rearranged as $\{G_1, G_2, \ldots, G_n\}, n \in \{1, 2, \ldots, N\}$, where $G_n = \{F^l_{t-n}, F^l_t, F^l_{t+n}\}$ is a downlink sequence from the previous F^l_{t-n} frame. The frame target is F^l_t, and the last frame is F^l_{t+n}. Note that the target conditions are shown for each group and this method can easily be generalized to any input frame. Clustering allows for the clear and efficient integration of adjacent images captured at different time intervals. First, it ensures that the contribution of adjacent images with varying time intervals is equally captured, particularly in images with significant distortion, mismatch, and motion blur. When an area is grouped (e.g., due to congestion), any lost information can be recovered by another group. This indicates that information from different clusters complements each other. Secondly, the reference conditions in each cluster guide the model in extracting useful insights from adjacent frames, enabling an efficient extraction and synthesis of information.

3.3. **Mixed spatial-temporal convolution (MSTC).** In Figure 4, we demonstrate the construction of an MSTC when given successive features from several frames with a size of $T \times H \times W \times C$ as input. T, H, W, and C denote the temporal duration, height, and width in the spatial domain and the number of channels. We begin by independently applying 2D spatial convolution of shared weights to each frame, as detailed below:

$$F^{2D}_t = C_{sc}(F^l_t), \tag{4}$$

where t is the temporal dimension index, F_t is the input feature of frame t, and $C_{SC}(\cdot)$ is the 2D spatial convolution process seen in Figure 5. F^{2D}_t depicts the t-the frame has

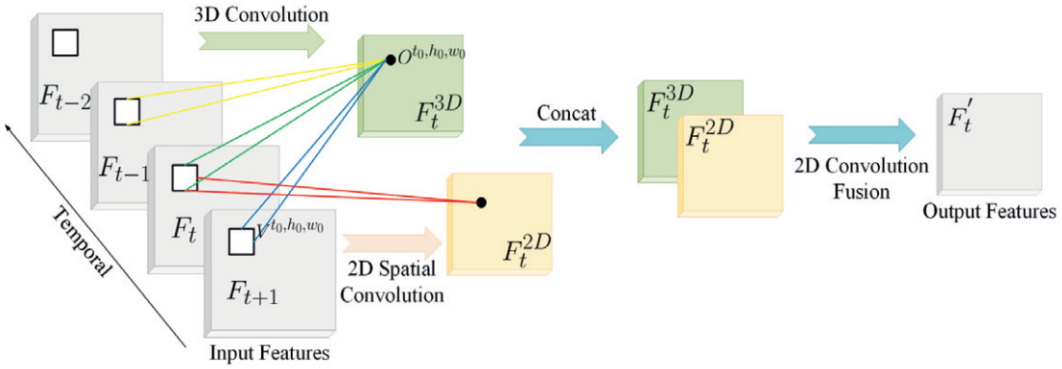

FIGURE 4. Illustration of a Mixed Spatial-Temporal Convolution (MSTC) that integrates 3D temporal information as supplement materials into the features generated by 2D spatial convolution

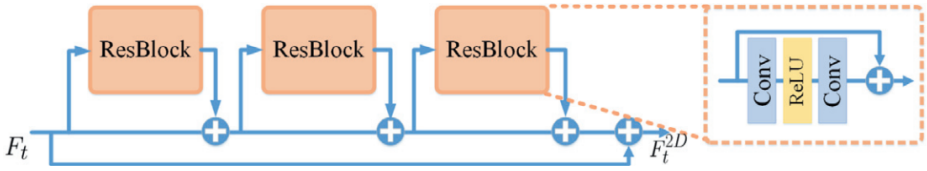

FIGURE 5. Illustration of 2D spatial convolution

extracted feature maps, which are self-independent features from each frame. Following the 2D spatial convolution procedure, the independently recovered features only include the learned spatial information from each frame.

We include a 3D convolution operation layer $C_{TC}(\cdot)$ to explore the temporal correlation between consecutive frames in addition to complementing the temporal signals in the video for more in-depth feature development. The kernels of a 3D convolution operation are represented as a 5D tensor. $\kappa \in R^{c_k \times n_k \times t_k \times h_k \times w_k}$, where c_k, t_k, h_k, w_k are the kernel sizes for the C, T, H, W dimensions, and n_k is the number of kernels. A 3D convolution layer accepts the input 3D features $\mathbf{F}_n^l = \{F_{t-n}^l, F_t^l, F_{t+n}^l\}$ and produces the feature maps F_t^{3D} by implementing convolution along the temporal dimensions of the inputs,

$$F_t^{3D} = \kappa \otimes \mathbf{F}_n^l \tag{5}$$

3.4. **Our TGA Procedure.** In Figure 1, it is shown that our TGA is modified in several parts. The main modifications involve registration using PDC for motion compensation and feature extraction using MSTC in one of the reconstruction stages. In addition, we also made dynamic depending on the temporal clustering stage according to the amount of frame input and reduced the repetition of skipped connections. Our TGA modification processes are summarized in the following Procedure 1.

4. **Experiments.** In order to assess the method, an experiment was carried out to compare the detection results on frames with super-resolution and without super-resolution. Additionally, the detection results will be compared to the actual plate location by matching the number of empirical-looking plates, i.e., by cutting the image of the detected plate to the size of the visible plate based on the detection results. The dataset we utilize is proprietary data collected directly at one-way highway intersections.

Procedure 1 TGA Modification with PDC registration and MSTC feature extraction

Input : LR image sequential data $\mathbf{X} = \{I_{t-n}, ..., I_t, ..., I_{t+n}\}$, number of groups N
Output : HR image of I_t

Procedure:

1. For each image in $\mathbf{X}$, extract a flat feature such that $\mathbf{F} = \{F^i_{t-n}, ..., F^i_t, ..., F^i_{t+n}\}$ by utilizing shared 3×3 convolutional layers $C_i(\cdot)$
2. Registration by applying PDC according to Equation (1)-(2) to obtain Z_{t-n} and acquire the feature map F^l_{t-n} on each element of $\mathbf{F}$ using the attention module in Equation (3).
3. Divide the registered data into groups of length N, i.e. $\{G_1, G_2, \ldots, G_n\}$, for $n \in \{1, 2, \ldots, N\}$ where $G_n = \{F^l_{t-n}, F^l_t, F^l_{t+n}\}$.
4. Perform the feature extraction process with MSTC according to Figure 4 with a sequence of processes using 2D spatial convolution and 3D convolution operations in Equations (4)-(5) to get the output feature F'_t
5. Apply attention map to get the upsampling sub-pixel image $\hat{F}'_t$ for each F'_t.
6. Fuse features from multiple temporal groups based on the attention map results with the Inter-group Fusion Module (IFM) into $\hat{F}_t = \text{IFM}(\{\hat{F}'_t\})$ and apply bicubic upsampling (BU) on the LR image separately, which is $\hat{I}_t = \text{BU}(I_t)$.
7. concatenate the feature result of $\hat{F}_t$ with the bicubic upsampling result of $\hat{I}_t$ to get the HR image of $I_t = \text{Concatenate}(\hat{F}_t, \hat{I}_t)$.

It is a 30 FPS video with a 480×640 resolution. The video contains 8342 frames, of which we divided 80% for training data and 20% for test data. On the training data, we applied a Gaussian blur with a standard deviation of 1.6 and performed four times downsampling to produce a low-resolution image. We used PSNR and SSIM scores to evaluate the performance of the model.

In the MSTC module, the space-time extraction runs once for each cluster and continues in the cluster-internal merge module. We used three 2D units for the spatial feature extractor, followed by 18 2D units with 3D complete blocks and 3D integration to integrate information in each group. In the intergroup compositing module, we used N convolutional 2D units in 2D cubes and set the channel size of the convolutional layer to 16 2D units. Our network accepts an odd number of low-resolution input frames. The model is controlled by pixel-by-pixel L1 loss and optimized by Adams Optimizer.

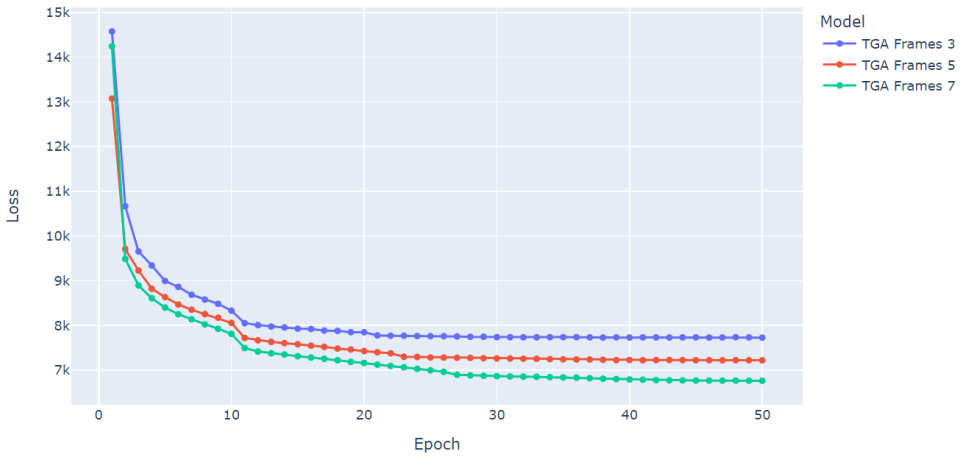

FIGURE 6. Loss Value for Each Epoch on the Our TGA Model

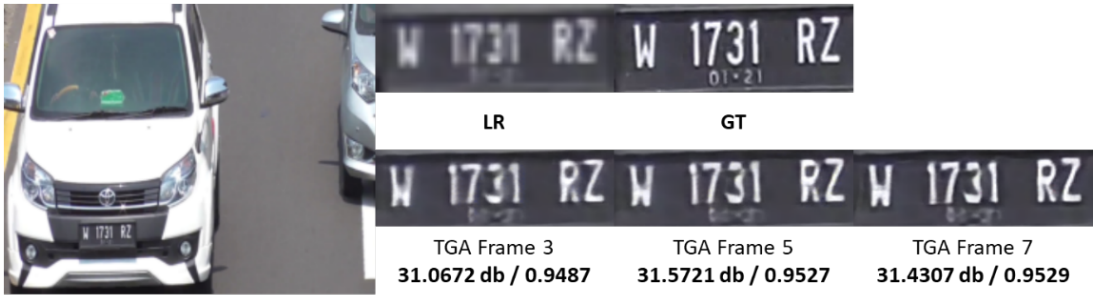

FIGURE 7. Qualitative comparison on video for 4× SR (PSNR / SSIM)

TABLE 1. Comparison of evaluation results of our TGA and original models

Model	Frames	PSNR	SSIM	Time (Seconds)
TGA	5	33.121333	0.965708	0.344177
	7	32.962441	0.964119	0.344963
Our TGA	5	**33.341816**	**0.966207**	**0.221901**
	7	33.291556	0.965986	0.568245

Weight loss during training is set at 5×10^{-4}, the learning rate is initially set at 2×10^{-3} and then reduced by 0.1 every ten epochs at 50. The mini-batch size is set to 64 and the training data is expanded by mirroring and twisting with a probability of 0.5. All tests are run on servers with Python 3.6.4, PyTorch 1.1, and Nvidia P100 GPUs.

In the training process, the loss value is calculated for each iteration, and then the values from several iterations are averaged to determine the loss for each epoch. Loss training results from each epoch are presented in Figure 6. In the figure, the loss value of each epoch is decreasing, meaning that the more epochs are done, the better the results. The best loss value is at the 50th epoch, so the model at that epoch is used for experimenting with the testing data. Table 1 presents the PSNR and SSIM results. It shows that our TGA model with input frame 5 has the largest PSNR and SSIM accuracy compared to other models, namely 33.341816 db and 0.966207. These results show that our modifications have been successful. Figure 7 shows a visual comparison of one of the frames in the validation data. This is evidenced by the SSIM value, which is close to one. It means that the super-resolution image has a good level of similarity with ground truth.

The comparison results of the number of license plates detected with our super-resolution method are shown in Table 2. From the table, the use of super-resolution improves detection accuracy. Our TGA with 5 and 7 frames gives a very significant increase in accuracy compared to no super-resolution. The average accuracy obtained for all frames was 90.37%, an increase of 4.23% compared to without Our TGA. These results are also in line with the precision and recall values, which provide positive differences at relatively the same time. In conclusion, super-resolution

TABLE 2. Comparison of detection results with our TGA

Model SR		Accuracy	Precision	Recall	Time (Seconds)
No		86.67%	88.64%	97.50%	1.334589
	Frames 3	88.89%	90.91%	97.56%	1.342622
Our TGA	Frames 5	91.11%	93.18%	97.62%	1.375788
	Frames 7	**91.11%**	**93.18%**	**97.62%**	1.388070

with Our TGA generally improves accuracy, precision, and recall. Therefore, this research contributes positively to the problem of vehicle license plate detection.

5. **Conclusion.** Our research proposes a new hierarchical neural network that implicitly integrates time information. The input sequence is rearranged into several contiguous groups with different frame rates to use complementary information between frames effectively. Clustering allows you to extract spatiotemporal information by combining 2D and 3D convolutions or hierarchically combining intergroup and intergroup fusion modules. The intra-group fusion module extracts features within each group, and the intra-group fusion module adaptively borrows complementary information from different groups. Furthermore, a high-speed spatial alignment technique has been suggested for addressing video footage featuring significant movements. The modification method can rebuild a high-quality, high-resolution framework and ensure consistency over extended periods. It demonstrates a license plate position accuracy of 90.37% without utilizing SR models and 86.67% using the Yolo V3 network during recognition testing. In the future, we expect to improve the performance of the TGA model so that it can be utilized in other fields.

REFERENCES

[1] C.-M. Kuo, C.-J. Hsu, Y.-X. Zheng, and H.-J. Ding, A novel key-frame extraction approach for semantic video processing., *J. Inf. Hiding Multim. Signal Process.*, vol. 14, no. 3, pp. 136–147, 2023.

[2] R. N. R. Wijaya, B. Setiyono, M. Yunus, and D. R. Sulistyaningrum, Operator-n layer construction for optimizing capsule network methods in image classification problems., *J. Inf. Hiding Multim. Signal Process.*, vol. 14, no. 3, pp. 90–101, 2023.

[3] L. P. Maguluri, J. Ananth, S. Hariram, C. Geetha, A. Bhaskar, and S. Boopathi, Smart vehicle-emissions monitoring system using internet of things (iot), in *Handbook of Research on Safe Disposal Methods of Municipal Solid Wastes for a Sustainable Environment*, pp. 191–211, IGI Global, 2023.

[4] G. Akbar, M. M. Iqbal, S. Ramzan, S. Majeed, and M. Farooq, License plate identification using machine learning techniques, *Journal of Computing & Biomedical Informatics*, 2024.

[5] H. Moussaoui, N. E. Akkad, M. Benslimane, W. El-Shafai, A. Baihan, C. Hewage, and R. S. Rathore, Enhancing automated vehicle identification by integrating yolo v8 and ocr techniques for high-precision license plate detection and recognition, *Scientific Reports*, vol. 14, no. 1, p. 14389, 2024.

[6] B. Setiyono, D. A. Amini, and D. R. Sulistyaningrum, Number plate recognition on vehicle using yolo-darknet, in *Journal of Physics: Conference Series*, vol. 1821, p. 012049, IOP Publishing, 2021.

[7] N. Haque, S. Islam, R. A. Tithy, and M. S. Uddin, Automatic bangla license plate recognition system for low-resolution images, in *2022 4th International Conference on Sustainable Technologies for Industry 4.0 (STI)*, pp. 1–6, IEEE, 2022.

[8] M. Hijji, A. Khan, M. M. Alwakeel, R. Harrabi, F. Aradah, F. A. Cheikh, M. Sajjad, and K. Muhammad, Intelligent image super-resolution for vehicle license plate in surveillance applications, *Mathematics*, vol. 11, no. 4, p. 892, 2023.

[9] B. Setiyono, M. Hariadi, and M. H. Purnomo, Survey of super resolution using phased based image matching, *Journal of Theoretical and Applied Information Technology*, vol. 43, no. 2, pp. 245–253, 2012.

[10] H. Ibrahim, O. M. Fahmy, and M. A. Elattar, License plate image analysis empowered by generative adversarial neural networks (gans), *IEEE Access*, vol. 10, pp. 30846–30857, 2022.

[11] N. O'Mahony, S. Campbell, A. Carvalho, S. Harapanahalli, G. V. Hernandez, L. Krpalkova, D. Riordan, and J. Walsh, Deep learning vs. traditional computer vision, in *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1*, pp. 128–144, Springer, 2020.

[12] X. Zhu, Z. Li, J. Lou, and Q. Shen, Video super-resolution based on a spatio-temporal matching network, *Pattern Recognition*, vol. 110, p. 107619, 2021.

[13] X. Wang, J. Chen, Z. Han, Q. Zhu, and W. Ruan, Real-time video deraining via global motion compensation and hybrid multi-scale temporal correlations, *IEEE Signal Processing Letters*, vol. 29, pp. 672–676, 2022.

[14] Y. Jo, S. W. Oh, J. Kang, and S. J. Kim, Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3224–3232, 2018.

[15] W. Sun, D. Gong, J. Q. Shi, A. van den Hengel, and Y. Zhang, Video super-resolution via mixed spatial-temporal convolution and selective fusion, *Pattern Recognition*, vol. 126, p. 108577, 2022.

[16] S. Ren, J. Li, T. Tu, Y. Peng, and J. Jiang, Towards efficient video detection object super-resolution with deep fusion network for public safety, *Security and Communication Networks*, vol. 2021, no. 1, p. 9999398, 2021.

[17] X. Wen and M. Zhaohui, Video super resolution enhancement based on two-stage 3d convolution, in *2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP)*, pp. 711–715, IEEE, 2021.

[18] H. Li and P. Zhang, Spatio-temporal fusion network for video super-resolution, in *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9, IEEE, 2021.

[19] Y. Zou, Y. Zhang, J. Yan, X. Jiang, T. Huang, H. Fan, and Z. Cui, License plate detection and recognition based on yolov3 and ilprnet, *Signal, Image and Video Processing*, vol. 16, no. 2, pp. 473–480, 2022.

[20] T. Isobe, S. Li, X. Jia, S. Yuan, G. Slabaugh, C. Xu, Y.-L. Li, S. Wang, and Q. Tian, Video super-resolution with temporal group attention, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8008–8017, 2020.

[21] Y. Tian, Y. Zhang, Y. Fu, and C. Xu, Tdan: Temporally-deformable alignment network for video super-resolution, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3360–3369, 2020.

[22] S. Qi, J. Du, M. Wu, H. Yi, L. Tang, T. Qian, and X. Wang, Underwater small target detection based on deformable convolutional pyramid, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2784–2788, IEEE, 2022.