Grenze International Journal of Engineering and Technology, Jan Issue

A User-Friendly Approach to Object Removal: CGANs and STTN for Enhanced Image and Video Editing

Preeti Bailke[1], Aditya Naikwad[2], Aman Pal[3], Aryan Naik[4], Bhairavi Deshmukh[5] and Tejas Bharambe[6]
[1]Associate Professor, Vishwakarma Institute of Technology, Pune, India
Email: preeti.bailke@vit.edu
[2-6]UG Students, Department of Artificial Intelligence and Data Science, Vishwakarma Institute of Technology, Pune, India
Email: {arun.aditya211, aman.pal21, aryan.naik21, bhairavi.deshmukh21, tejas.bharambe21}@vit.edu

***Abstract*— This research introduces an advanced object removal tool addressing the challenges in computer vision and image processing. Leveraging Generative Adversarial Networks (GANs) with Python, TensorFlow, and OpenCV, the tool utilizes conditional GANs for efficient learning of patterns and structures. By employing matched image sets comprising original images and corresponding object masks, the program generates realistic and aesthetically pleasing in-painted images. Additionally, a user-friendly web application is developed, allowing users to select input images and define specific regions for removal. The tool's versatility extends to video editing, incorporating a Spatial-Temporal Transformer Network (STTN) for seamless removal of objects in both spatial and temporal dimensions. This comprehensive solution enhances user engagement and expands the tool's applicability to eliminating unwanted elements from videos.**

***Index Terms*— Object removal, conditional GANs, image inpainting, computer vision.**

I. INTRODUCTION

It is sometimes important to eliminate undesired things from photos. For example, the inclusion of undesirable objects in old photographs or damaged images might lower the final product's quality. By removing these hindrances, object removal tools make restoration efforts possible and enhance the image's aesthetic appeal and structural integrity. People frequently want to add their unique stamp to photographs by removing unwanted or distracting elements. In order to meet requirements or create compositions that are aesthetically pleasing, it is frequently necessary to remove objects from images in the fields of graphic design, advertising, or content creation. An object removal tool can aid in the obfuscation or blurring of such content in situations where private or sensitive information needs to be protected. This is especially important when posting pictures online or in public. Removing unnecessary or obstructive objects from crime scene photographs can improve the clarity of the evidence and help with analysis in forensic investigations. Unwanted elements can change the overall caliber and consistency of a collection of images, such as sensor dust spots or lens flares.

By removing these discrepancies across multiple images, an object removal tool helps to maintain visual coherence. The removal of objects or persons in a video too can be of significant importance in various contexts. Like, in order to keep the narrative in check and coherent in a video, objects or people can be eliminated.

Sometimes, certain aspects of a scene may be out of sync with the intended plot or unintentionally reveal unwanted information. The narrative of the video can be preserved or altered as desired by removing these objects or people,

Grenze ID: 01.GIJET.10.1.211

resulting in a seamless and interesting storytelling experience. There can be circumstances where it might be necessary to remove identifiable people or sensitive information from a video in order to protect privacy or secure consent. This is particularly important when dealing with footage taken in public areas, crowds, or recordings that include individuals whose inclusion has not received their express consent.
Removing an item or a person can help protect privacy rights and uphold moral and legal obligations. The removal of objects or people can help isolate particular events or make important details more visible in forensic analysis and surveillance scenarios. Investigators or analysts can concentrate on particular aspects of the video to glean important information or evidence by removing irrelevant components. An unwanted person or object can be eliminated from a video to improve its aesthetic appeal. By removing distractions and sharpening compositions, it enhances the visual appeal of the content as a whole. This can be especially important in artistic or professional video production, where it's important to deliver an engaging and aesthetically pleasing experience.
In this research, a web application having an image as well as video editor has been created. Unwanted things in an image can be deleted using the image editor, which is built with Python, TensorFlow, and OpenCV and is based on conditional Generative Adversarial Network (CGAN) idea. Unwanted persons can be eliminated from a video using the video editor. For video inpainting, it employs the concept of a Spatial-Temporal Transformer Network.
The research described in this paper outlines a user-friendly approach to object removal, leveraging the power of conditional Generative Adversarial Networks (CGANs) and Spatial-Temporal Transformer Networks (STTN) for advanced image and video editing. To achieve our objectives, we employ a combination of technologies, including Python, TensorFlow, and OpenCV. For image editing, our system employs CGANs, where the core components include an autoencoder and a discriminator. In the realm of video editing, we base our approach on STTN, a framework that efficiently addresses object removal across both spatial and temporal dimensions. To improve user engagement, we've developed a web application that allows users to interact with our tool. Users can select an input image and specify the region they wish to eliminate, or they can apply the system to remove unwanted individuals from videos. By integrating these techniques, our research aims to provide a comprehensive solution for object removal, offering users a simple and effective way to enhance the aesthetic appeal and usability of their visual content.

II. Literature Survey

Sourav Kulkarni in [1] discussed some basic statistical algorithms and techniques for removing unwanted objects. The techniques in this paper, which require less time and effort to accurately reconstruct the image than Image inpainting, also known as object removal, and manual photo correction, encompass various techniques used in computer vision and image processing. Straight forward yet very effective at removing unwanted objects from images.
Antonio Criminisi et al [2] presented an algorithm for the removal of large objects and their replacement with aesthetically pleasing backgrounds in digital images. This study presents a novel method for image inpainting that overcomes some of the drawbacks of earlier techniques. The method was also able to handle complex textures and structures unlike the previous ones.
Z.-Q. Zhao et al [3] investigates the Deep Learning based object identification framework, examines their innovations, and emphasizes the advantages of utilizing deep neural networks for reliable and precise object detection. It gives a summary of the present state of the art, illuminating the advancements made in this area.
Rosana Balanescu et al [4] presented two algorithms that combine and search for missing information in nearby frames to A technique that involves taking many photographs of the scene, breaking them into macroblocks, matching corresponding macroblocks across the images, and blending them together to create a single, corrected image can be used to remove unwanted elements from a series of subsequent images.
An inventive method for object removal utilizing exemplar-based inpainting in photos was put out by Criminisi et al. [5]. It uses similar or extracted exemplars from the image to fill in the gaps left by the removal of certain areas. With this technique, smooth object removal and visual coherence are both guaranteed. Their contribution resolves obstacles to object removal in a variety of applications. This method produces realistic and seamless results and can handle complex structures and textures.
A thorough explanation of the Kriging Interpolation Technique, a diffusion-based object removal technique was given in [6]. Using this technique, even if the image's texture is different from the unwanted object's, the unwanted object can still be removed. The method performs better with smaller block sizes, producing results with low MSE and high PSNR. Marcelo Bertalmio at al [7] Presents an innovative technique for digitally restoring missing areas in still images, designed to replicate the methods used by expert restorers. This algorithm automatically replaces

the selected regions with surrounding information, chosen by the user. It intelligently fills in the isophote lines that approach the boundaries of the regions during the restoration process. Unlike previous approaches, this method eliminates the need for the user to indicate the source of the new information.
Chang, Ya-Liang, et al [8] A new model for video inpainting using deep learning was introduced, which focuses on filling in missing areas in videos with free-form masks. To enhance temporal consistency, a novel Temporal PatchGAN loss was introduced. The model also addressed the uncertainty associated with free-form masks by incorporating 3D gated convolutions. In order to train and evaluate video inpainting models, a dataset called free-form video inpainting (FVI) was created, consisting of videos and a newly developed algorithm for generating free-form masks.
Xu, Zongben et al. [9] developed an example-based inpainting method by analyzing the differences between image paths. Anat Levin et al [10] used statistical learning to investigate the issue of learning to paint utilizing the remaining portion of the image. The histograms of local characteristics, the distribution, and a training image are used to create an exponential family distribution over images. Then, using this image-specific distribution to fill the hole, the most expected image given the boundary and its distribution is found. Loopy belief propagation is used to carry out the optimizations.
Zeng et al [11] proposed learning a Spatial-Temporal Transformer Network which would be used in video inpainting in this paper. It was proposed in order to optimize STTN using a spatiotemporal adversarial loss while filling empty regions in all the given input frames at the same time utilizing self-attention. Granados, Miguel, and colleagues [12] present a novel method for finishing films that can manage complex sequences with dynamic backgrounds and erratic moving objects. The idea that a removed object's spatiotemporal hole can be filled can be expanded by using data from other areas of the movie when the occluded objects were visible.
N Sasipriyaa et al. [13] proposed a new deep learning-based method for HCR on a dataset of handwritten Tamil characters. Handwritten character recognition (HCR) is a challenging task due to the variability of human handwriting.This method is based on a combination of conditional generative adversarial networks (cGANs) and CNNs.In the case of HCR, cGANs can be used to generate images of handwritten Tamil characters. CNNs can then be used to classify these images. Chuan Qin et al. [14] addresses the need for robust image watermarking, presenting classical and deep learning-based methods. It introduces a novel attack approach, Conditional Generative Adversarial Nets (CGAN)-based Attack on Deep Robust Watermarking (CADW), designed to undermine deep watermarking models effectively. CADW, transferable across different networks, achieves high success rates in attacking StegaStamp and HiDDeN, making it a versatile standard for evaluating watermarking robustness.
Rafael Molossi Escher et al. [15] focused on reducing attention-based state-of-the-art video inpainting's computational complexity, increasing its performance, and making it easier to utilise on low-end GPUs. Fast Spatio-Temporal Transformer Network (FastSTTN), an extension of the Spatio-Temporal Transformer Network (STTN) that maintains state-of-the-art video inpainting accuracy while reducing memory usage up to 7 times and execution time by roughly 2.2 times thanks to the adoption of Reversible Layers was introduced in this work. Yulai Xie et al. [16] addresses the challenge of predicting urban crowd flow for traffic management and city-level studies. It introduces the MultiSize Patched Spatial-Temporal Transformer Network (MSP-STTN) for short- and long-term grid-based crowd flow prediction. MSP-STTN employs self-attention and cross-attention mechanisms, categorized space-time expectation, and auxiliary tasks to enhance prediction accuracy. The proposed model is evaluated on real-world datasets, achieving competitive results for short-term and practical long-term crowd flow prediction.

III. IMPLEMENTATION

1. Image Editor

The web application has 2 parts – Image Editor and Video Editor.
Firstly, the image editor has been discussed here. Here conditional Generative Adversarial Network (CGAN) has been used for deep image completion after removing the objects from it. This network architecture consists of 2 parts: the autoencoder and the discriminator. The given diagram symbolizes the flow of working of conditional GANs.
The generative network or autoencoder consists of semantic feature extractor and a simple generator. Inputting a corrupted image 1, the autoencoder attempts to reconstruct a clean copy of the image. Semantic feature extractor, also known as encoder, extracts high level semantic features such as shapes, objects, textures, or other relevant characteristics from the input image and compresses it into lower-dimensional representation. The generator takes

this compressed representation and reconstructs an output that is as close as possible to the input image, minimizing the information loss incurred during the compression process. The autoencoder is able to recognize and replicate the key features of the input image because the generator and semantic feature ex-tractor team up to learn an efficient representation and reconstruction process.

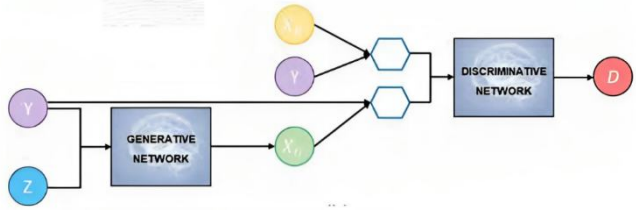

Fig1. Working of CGANs

To train this autoencoder, the discriminative network or discriminator is used. After that, it has a number of convolutional layers, a fully connected layer and global average pooling. It is taught to distinguish between actual and fake images, providing a signal to the autoencoder to generate more realistic and aesthetically accurate images. The adversarial training procedure comprises training the generating and discriminative networks in opposing directions. During training, the discrimination and generating networks are iteratively updated. While the discriminator seeks to increase the accuracy of its discrimination, the generative network aims to reduce the discriminator's ability to distinguish between real and completed images. These competitive dynamic forces the generative network to produce images that are statistically accurate and visually appealing.

2. Video Editor

For video inpainting an STTN has been employed. It is divided into three sections: a frame-level encode, a frame-level decoder and spatial-temporal transformers. Below Figure 2 explains the overall flow of STNN.

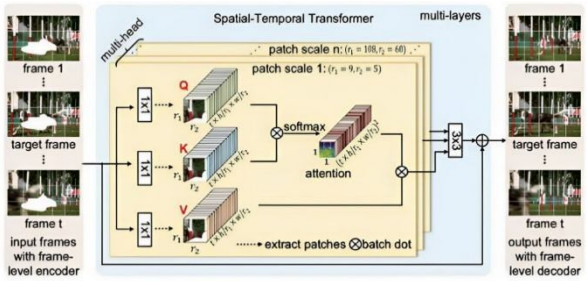

Fig2. Overall flow of STTN

Frame-level encoder takes frames of a video sequence as input. Each frame's spatial data is encoded, and useful features that capture the content and context of the video frames are extracted. To learn these spatial features, it uses Convolutional Neural Networks (CNNs). The spatial-temporal transformers in the STTN are designed to simultaneously record both spatial and temporal information in the context of video inpainting. This component is the core of the STTN and is responsible for performing spatial-temporal transformations They pay attention to the spatial and temporal contexts of each frame, taking into account the connections between various regions both within the current frame and across subsequent frames. This aids the model's comprehension of the video's content and temporal coherence, which helps it produce accurate and coherent in-paintings. Frame-level decoders aim to convert characteristics into appropriate frames. It reconstructs the in-painted frames using the learned spatial and temporal representations. Convolutional or transposed convolutional layers are frequently used by the frame-level decoder to upsample and produce the final in-painted frames.

Then the T-PatchGAN has been used as the discriminator network for adversarial training. A PatchGAN is a discriminator network that works on local image patches rather than the entire image when used for video inpainting. Instead of classifying the realism of the entire image, it does so for individual patches, enabling more precise evaluation. With this strategy, the generator network is encouraged to produce local patch-level results that are realistic and visually consistent.

The "T" in T-PatchGAN stands for "temporal," denoting that the algorithm is made to take into account the temporal information contained in video sequences. When determining how realistic the generated video in-

paintings are, the T-PatchGAN discriminator takes both the spatial and temporal contexts into account. The generated content is checked to make sure it fits well within the temporal context of the video by examining the consistency and coherency of local patches across adjacent frames. The generator network is trained to produce video in-paintings that exhibit both temporal coherence with the neighboring frames and visual plausibility at the local patch level using the T-PatchGAN as the discriminator. This makes it easier to create in-painted videos with natural-looking spatial and temporal content that are visually consistent. Utilizing a combination of loss functions, optimization is used during the training process. Adversarial loss, perceptual loss, and reconstruction loss are some of these loss functions.

IV. Results And Discussion

Here are some screenshots of the web application created using Streamlit. Figure 3 denotes the first page having a small introduction of the application and navigation bar at the left side to go to other pages.

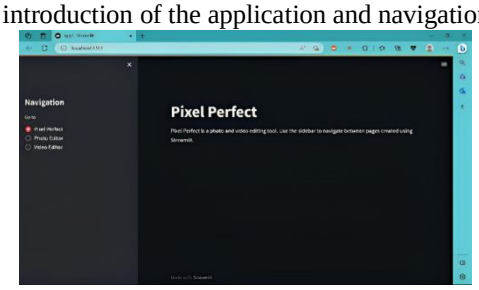

Fig3. First page of website

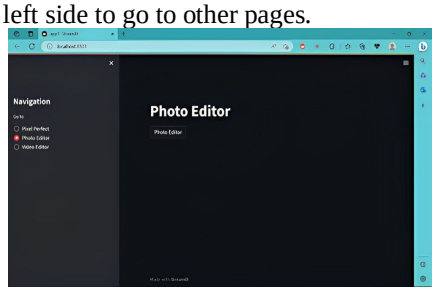

Fig4. The second page of website

Figure 4 shows the second page, i.e., of image/photo editor. It consists of a button named Photo Editor which on getting clicked shifts to another page where we can perform image editing. Figure 5 shows an example of it. The unaltered original image is depicted on the left side of the illustration, providing an untainted portrayal of the visual content under evaluation. Using a pointer device, users can highlight any exact place inside this image. When the 'Enter' key is pressed, the image displayed on the right side is modified, with the selected area eliminated. This interactive functionality not only allows the user to focus on certain areas of the image, but it also allows for dynamic, real-time editing to improve the material based on their preferences. This ground-breaking feature has enormous potential for use in research, visual content analysis, and image alteration, providing a versatile tool for selective highlighting and editing.

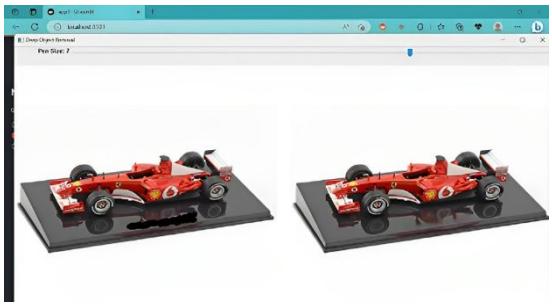

Fig5. Image editor window

Figures 6 and 7 depict the third page of the interface, which is an important component for video editing. This page allows users to engage in extensive video editing by requesting important user inputs, especially the movie name and the matching mask. Users can start the editing process by entering these settings, allowing them to make modifications to the original video content.

Furthermore, the user interface provides straightforward capability for video playback, where the original video, as well as the altered version, may be played simply clicking on the selected but-ton. This function enables users to visually examine and compare the differences between the original and edited video content, making it a useful tool for quality control and visual analysis.

Figures 6 and 7 combine the use of two significant picture quality assessment metrics, namely Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM), to further enhance this platform's analytical

capabilities. These metrics are easily computed by the user with a single click on the appropriate buttons, providing quantitative measures of the video's quality and re-semblance to the original source material. This multimodal approach to video editing and evaluation not only improves the user experience, but also highlights the system's potential utility for research and multimedia content alteration.

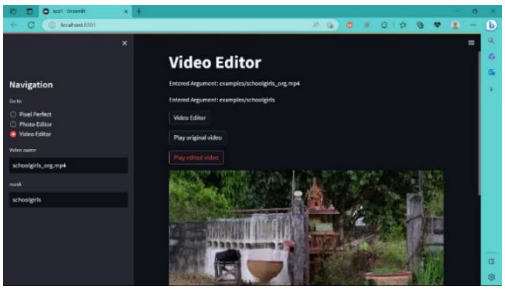

Fig6. Playing the edited video

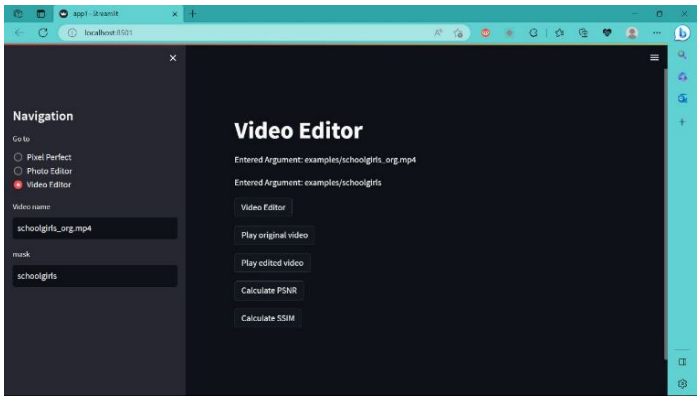

Fig7. The third page of website

By calculating the MSE (mean squared error) between corresponding frames, PSNR evaluates the generated video's quality in comparison to the original. PSNR values over 30 dB are regarded as good while values over 40 dB as excellent. The resultant PSNR value was 34.96dB which is considered between good to excellent according to standards.

V. Conclusion

The current research study has established a solid foundation for object removal in photos and videos utilizing CGANs and STTN. These technologies have the ability to address a wide range of real-world and creative use cases, including privacy protection, aesthetic enhancement, and video storytelling. The future scope of study spans multiple fascinating directions, ranging from real-time processing to ethical considerations, and it offers the possibility of significantly increasing the usability and accessibility of object removal tools across various domains. This study is a valuable contribution to the field of computer vision and image processing, with the potential to impact a wide range of applications.

Future Scope

The current research project has demonstrated the successful development of an object removal tool using conditional Generative Adversarial Networks (CGANs) for image inpainting and a Spatial-Temporal Transformer Network (STTN) for video editing. However, several avenues for future research and expansion can be identified, offering exciting opportunities for further exploration and innovation:

- *Object Replacement Tool*: One critical direction for future research is the development of an object replacement tool. This tool would allow users to not only remove unwanted objects but also seamlessly replace them with desired elements. Research in this area can explore advanced algorithms and machine learning techniques for accurate and contextually relevant object replacement.

- *Object Addition Functionality*: Expanding the tool's capabilities to include object addition functionality is a key research avenue. Enabling users to insert objects into images with ease would cater to applications in graphic design, advertising, and creative content generation. This feature could enhance the tool's versatility and practicality.
- *Process Time Optimization*: Future research can focus on reducing processing time. This can involve optimizing algorithms, leveraging parallel processing, or exploring hardware acceleration to achieve real-time or near-real-time performance for object removal and re-placement.
- *Enhanced Accuracy*: Improving the accuracy of the object removal and replacement process is paramount. Researchers can delve into refining machine learning models, exploring more extensive training datasets, and enhancing the object recognition and inpainting capabilities to ensure the highest level of precision.
- *Real-time Object Tracking:* Investigating real-time object tracking capabilities can enhance the tool's performance in dynamic video content. This could include developing methods for tracking objects as they move within a video, ensuring consistent and accurate object removal and addition.

REFERENCES

[1] S. Kulkarni, "REMOVAL OF UNWANTED OBJECTS FROM IMAGES USING STATISTICS," ICTACT Journal on Image and Video Processing, vol. 9, no. 2, pp. 1887–1893, Nov. 2018.

[2] A. Criminisi, P. R. Perez, and K. Toyama, "Region Filling and Object Removal by Exemplar-Based Image Inpainting," IEEE Transactions on Image Processing, vol. 13, no. 9, pp. 1200–1212, Sep. 2004.

[3] Z.-Q. Zhao, P. Zheng, S. Xu, and X. Wu, "Object Detection with Deep Learning: A Review," IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

[4] R. Balanescu, A. Sterca, and I. Badarinza, Removal of Unwanted Objects from Still Photographs. 2020.

[5] A. Criminisi, P. R. Perez, and K. Toyama, Object removal by exemplar-based inpainting. 2003.

[6] L. Bangaru and V. K. Gupta, Object removal by Kriging Interpolation Technique. 2015.

[7] M. Bertalmío, G. Sapiro, V. Caselles, and C. Ballester, Image inpainting. 2000.

[8] Y.-L. Chang, Z. Liu, K.-Y. Lee, and W. H. Hsu, Free-Form Video Inpainting With 3D Gated Convolution and Temporal PatchGAN. 2019.

[9] Z. Xu and J. Sun, "Image Inpainting by Patch Propagation Using Patch Sparsity," IEEE Transactions on Image Processing, vol. 19, no. 5, pp. 1153–1165, May 2010.

[10] Levin, Zomet, and Weiss, Learning how to inpaint from global image statistics. 2003.

[11] Y.-H. Zeng, J. Fu, and H. Chao, "Learning Joint Spatial-Temporal Transformations for Video Inpainting," in Lecture Notes in Computer Science, Springer Science+Business Media, 2020, pp. 528–543.

[12] M. Granados, J. Tompkin, K. W. Kim, O. Grau, J. Kautz, and C. Theobalt, "How Not to Be Seen - Object Removal from Videos of Crowded Scenes," Computer Graphics Forum, vol. 31, no. 2pt1, pp. 219–228, May 2012.

[13] N. Sasipriyaa, P. Natesan, R. Anand, P. Arvindkumar, R. S. A. Prakadis, and K. A. Surya, "Recognizing Handwritten Offline Tamil Character by using cGAN & CNN," 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Apr. 2022, doi: 10.1109/icscds53736.2022.9760808.

[14] C. Qin, S. Gao, X. Zhang, and G. Feng, "CADW: CGAN-Based Attack on Deep Robust Image Watermarking," IEEE MultiMedia, vol. 30, no. 1, pp. 28–35, Jan. 2023, doi: 10.1109/mmul.2022.3213004.

[15] R. Escher, R. De, and P. Drews, "Fast Spatial-Temporal Transformer Network," SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Oct. 2021, doi: 10.1109/sibgrapi54419.2021.00018.

[16] Y. Xie, J. Niu, Y. Zhang, and F. Ren, "Multisize patched Spatial-Temporal Transformer Network for Short- and Long-Term Crowd flow prediction," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 11, pp. 21548–21568, Nov. 2022, doi: 10.1109/tits.2022.3186707.