JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Enhancing near duplicate image detection using spatial transformer networks in deep learning

[1]Varun S, Dr. [2]Nagaraja S R
[1,2]Department of CSE
Presidency University, Bangalore, India

Abstract: The research implements Spatial Transformer Networks (STNs) as an advancement strategy for near-duplicate image detection algorithms built with deep learning models. The Spatial Transformer Network (STN) helps the system master spatial transformation corrections for translation and rotation and scaling and cropping of images which are typical near-duplicate image variations. By integrating STNs into Convolutional Neural Networks (CNNs) together with Vision Transformers (ViTs) the models obtain the capability to normalize image alignment prior to feature extraction. Using STNs in image detection systems creates improved performance levels in near-duplicate image recognition. The investigation combines different STN architectures and integration methods to obtain optimal system performance results. This project enhances image retrieval methods and similarity analysis through the correction of SIFT and SURF limitations in addition to resolving current deep learning technique challenges. The experimental process will demonstrate the operational effectiveness of STNs for illuminating spatial differences in near-duplicate image detection

IndexTerms **- Near-duplicate image detection, Spatial Transformer Networks (STNs), Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), spatial transformations, image alignment, deep learning**

I. INTRODUCTION

The increasing problem of digital image duplication during the information age brings consequences for several business sectors starting with social media and extending to intellectual property security. Different professional fields require solutions to detect images with minor spatial variations caused by translation or rotation or scaling or cropping [1]. When applied to digital repository management near-duplicate detection aids both in improving organization practices and in fighting duplication and protecting intellectual assets. SIFT (Scale-Invariant Feature Transform), SURF (Speeded-Up Robust Features) together with ORB (Oriented FAST and Rotated BRIEF) operate as traditional detectors for image near-duplicates because they conduct feature detection and matching functions. The detection methods excel at small-scale data sets but experience issues when facing extensive spatial modifications and complicated styles according to [2].

The deep learning revolution through convolutional neural networks (CNNs) brought automatic feature detection automation while increasing classification accuracy and detection precision [3]. Four detection methods based on CNN struggle with severe spatial image modifications that include cropping and rotation during near-duplicate detection tasks. These limitations highlight the Analysts need upgraded approaches that handle spatial modifications throughout calculations yet maintain performance speed [4].

STNs (Spatial Transformer Networks) serve as an effective solution for managing this problem. Training STNs builds spatial transformation abilities that make the model automatically normalize and align input image data. STNs introduced to deep learning frameworks let the systems handle spatial changes to achieve better system accuracy and performance. The solution makes detection of near-duplicate images with prevalent transformations possible because of its effectiveness [5].

Several distinct elements comprise the proposed work which unites STNs with CNNs and Vision Transformers (ViTs). CNN architecture extracts image features effectively through its hierarchical structure but Vision Transformers succeed best at establishing worldwide picture relationships. Through integrating STNs the system executes preprocessing that corrects image spatial distortions to improve both extraction and comparison operations [6].

The research investigates various STN structural and configuration setups to identify their best possible integration methods. Experimental tests will verify what impacts various configuration options generate on the accuracy of systems while measuring their ability to detect near-duplicates. Further testing based on traditional detection methods and leading deep learning architectures will show the advantages that result from this developed methodology [7].

Literature evidence indicates that combining transformers with CNNs results in superior performance when applied to spatial transformation tasks. Research findings show that hierarchical feature fusion and attention mechanisms produce effective results for controlling spatial complexities of object detection and tracking operations [8][9]. Transformers working with local feature extraction methods of CNNs establish robust capabilities which allow the development of improved near-duplicate detection systems [10].

Multiple studies in the literature analyze spatial transformer and transformer technology for improving underwater image quality and detecting objects within videos and recognizing human faces in both domains. Modern neural networks display their prowess by transferring processed image information from one operation to another for achieving analysis goals. The utilization of spatial perception with time-based perception within transformers produces improved detection outcomes in challenging operating situations [14]. Spatial transformer layers implemented in multimodal fusion tasks help to boost their alignment performance while improving their overall robustness [15].

This study addresses current limitations of near-duplicate image detection by uniting STNs along with their capabilities while combining them with CNNs and ViTs. The research project aims to advance current image retrieval knowledge by creating accurate detection systems that produce real-life solutions.

II. PROPOSED DETECTION METHOD

The following subsection details the creation process of STN-based near-duplicate image detection. The designed framework uses Spatial Transformer Networks to process deep learning structures for managing image spatial variations across translation, rotation and scaling alongside cropping effects. The methodology progresses through four distinct phases namely data preprocessing before moving onto model architecture design for training sequence and evaluation step and optimization stage.

A. Data Preprocessing

The system requires data preprocessing as an essential preprocessing phase to make the dataset suitable for robust and generalizable use by incorporating diverse spatial variations. The initial process of dataset collection focuses on obtaining pairs of images which have been labeled as near-duplicates or non-duplicates. The COCO and ImageNet datasets serve as appropriate sources because they contain numerous examples of different image types. The currently available datasets fall short of representing all the image transformations which would be found in true-world near-duplicate situations. The implementation of augmentation techniques applies to perform virtual spatial modifications by randomly applying rotations and scaling and cropping and translatable transformations. The augmentations generate datasets that show real-world variation in order to enable effective model performance on various near-duplicate examples [1][2].

The preprocessing step of normalization serves to adjust the dataset for standardizing inputs into the model. The normalization process standardizes images through scaling to the values between 0 and 1 or calculates normalization based on mean and standard deviation of the dataset. Model training consistency becomes achievable through this step because it reduces the occurrence of numerical instability and speeds up the convergence process. Deep learning models need uniform input dimensions thus all images undergo a fixed resolution adjustment for efficient computation. The data resizing procedure maintains image spatial relationships when applied to fit model architectural requirements [3].

The final step in data preprocessing requires dividing the dataset into training and validation as well as testing components with 70:15:15 proportions. The division of data between training and evaluation purposes helps maintain objective measures for model performance assessment before and after training. The model optimization process takes place with the training set although the validation set helps choose parameters and the test set shows how well the model performs on new data. The correct partitioning of data decreases overfitting risk so the real-world system becomes robust. The near-duplicate image detection system will achieve both accuracy and scalability through this method [4].

B. Model Architecture Design

The model incorporates Spatial Transformer Networks (STNs) with deep learning frameworks because these produce richer results when dealing with spatial changes which include rotation scaling translation and cropping. STNs operate as preprocessing components through automatic feature transformation to achieve spatial consistency before the extraction process [5]. By using this preprocessing capability the system can manage various image variations and achieve better detection results for near-duplicate images because the feature extraction functions consistently. STNs create aligned spatial representations that let the following layers detect real features without spatial misalignments or spatial distortion interfering with their operation [6].

The STN framework contains three essential elements that make its operation possible. During the localization process the network determines transformation parameters including an affine matrix to specify how the input image requires spatial modifications [7]. During training the system learns these parameters which will help it handle diverse input scenarios for enhanced accuracy in alignment tasks. After transformation parameter predictions the grid generator develops sampling grids which function as pixel mapping directions from input to output [8]. Through the sampling process the input image undergoes spatial transformation-based normalization that produces an image ready for robust feature extraction [9]. The preprocessing operation plays a vital role in strengthening both the precision and stability of near-duplicate image detection systems according to [10].

Enhancing STNs with Hybrid Models

The research method suggests adding STNs to hybrid architectures to build better near-duplicate image detection solutions with improved accuracy and robust performance. Through spatial normalization STNs provide an effective mechanism that learns transformations to align input images before they reach downstream models. STN performance becomes stronger through

cooperation with sophisticated deep learning concepts. The combination of multiple architectures through a hybrid model allows detection of superior near-duplicate images by handling diverse spatial and global variations [5].

STNs form the initial element of the integrated system because they serve as an input processing component to synchronize images and normalize spatial deformations stemming from image transformations including rotation along with scaling and cropping. The CNNs perform feature extraction on processed outputs that result from alignment operations. The local feature extraction ability of CNNs focuses on textures and edges while performing near-duplicate image identification tasks. Feature representation becomes robust through CNNs because they specialize in processing specific image areas [8].

Vision Transformers (ViTs) join CNNs within the architecture to provide models that evaluate global relationships and spatial dependencies while maintaining CNNs' localized viewpoint. The entire picture analysis of ViTs relies on attention mechanisms that detect spatial connections between distant image parts. The feature allows detection of global variations and patterns that stand outside the awareness of CNN-based architectures. The combination of this hybrid system architecture makes sure detection occurs with complete awareness of detailed local information while maintaining international contextual understanding [11].

Through the combination of STN spatial alignment power with CNNs and ViTs the system develops complete knowledge of image modifications. The system effectively deals with difficult images containing major transformations and background clutter or multiple objects in the scene. The combination of detection approaches improves system accuracy and produces better generality across various datasets for real-life applications. The modular design of this architecture enables expansion and system improvements which keeping it adaptable to changing image detection needs [14].

- **STN + CNN + Vision Transformer (ViT)**

The Spatial Transformer Networks (sTN) function to adjust spatial image regions that handle local variances like cropping and scaling as well as image rotation to enable the model's focus on crucial image components despite such transformations. The experts in [11] demonstrate that CNNs extract detailed image features from spatial areas using multiple layers for sequential attribute extraction.[10]Vision Transformers (ViTs) process images through a distinct method that models all global connections inside images across the whole visual field. The network analyzes overall image spatial relationships to gain complete contextual information thus making it suitable for tasks that need an integrated view of visual data.

The detection performance receives an improvement from precise alignment which is enabled through robust feature extraction.

Pseudocode for the Proposed Integration:

```
Input: Near-duplicate images
Step 1: Spatial alignment using STN
Step 2: Extract local features using CNN
Step 3: Feed features to Vision Transformer for global
context modeling
Step 4: Compare feature embeddings to detect duplicates
Output: Classification of near-duplicate pairs
```

The adoption of attention mechanisms into Spatial Transformer Networks (STNs) provides improved duplicate image detection because they enable precise regional dynamics which decreases weaker area impact. Regional attention boosts the model's performance for detecting near-duplicate images because it establishes sharp separation of subtle image differences in similar inputs. The network can process image relations between different parts through the multi-head self-attention mechanism which splits features into multiple attention pathways. The processing of multiple image patterns simultaneously within various parts enables the network to achieve better image understanding. This approach proves beneficial for intricate visual material because it detects common patterns that exist between different regions of an image [11][12].

Twice-gated spatial refinement in the system produces essential background weighting to correct image information value based on its importance level. This selective weighting strategy enhances crucial image areas for analysis purposes while it reduces background interferences. Through combined functioning these mechanisms help STNs find precise distinctions between features that leads to higher accuracy rates and operational effectiveness and system stability. The system functions at peak performance levels for real-world image detection tasks due to enhancements which resolve issues related to lighting variation alongside orientation problems and background clutter conditions [13][14][15].

Pseudocode:

```
Input: Image pair (I1, I2)
Step 1: Apply Spatial Transformer Network (STN) for alignment
Step 2: Pass aligned features through Attention Layer
   - Compute attention weights for key regions
   - Focus on relevant features
Step 3: Extract features and compare embeddings
Output: Near-duplicate classification
```

The system's operation heavily depends on Convolutional Neural Networks (CNNs) because they function as the essential feature extractor that uses a hierarchical design to retrieve specific features from Spatial Transformer Network (STN) delivered normalized images. The local patterns identified by CNNs refer to edge detection together with texture recognition and shape analysis which serve crucially for near-duplicate image comparison [13]. Due to their widespread success in multiple image tasks ResNet and VGGNet remain the top choices for backbone networks because they perform excellently in producing Discriminative and Robust features. Such architectures develop abstract features through several convolutional and pooling layers which make the model focus on progressively higher-level patterns [4]. The model performs consistently efficient feature extraction operations throughout all types of challenging environmental conditions including various light conditions and equipment orientation and image quality specifications. The extracted features are subjected to processing through a fully connected layer which determines whether the image pairs match or do not match. The final output comes from the similarity evaluation of feature vectors where cosine similarity or Euclidean distance perform calculations for final classification [3].

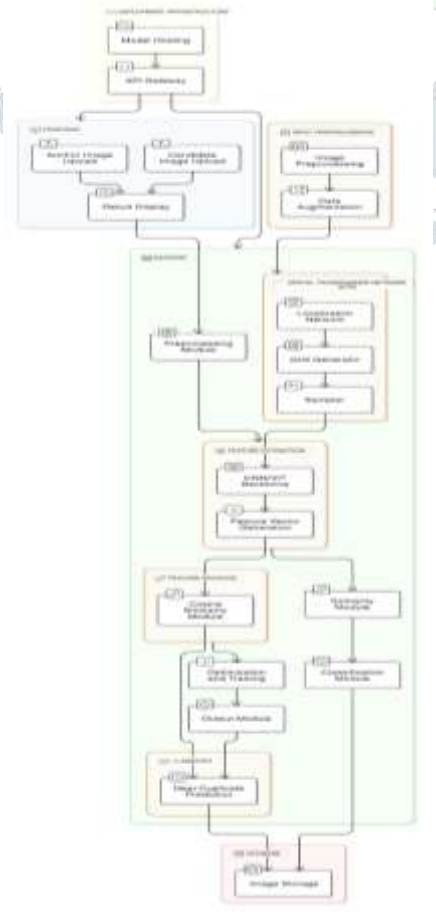

Figure 1: System Architecture

Vision Transformers (ViTs) serve as a replacement alternative to CNNs for feature extraction because they demonstrate superior capability in understanding global dependencies found across image features. The self-attention of ViTs surpasses CNNs' local pattern recognition through their ability to study all image relationships across the entire image. The system provides high value when analyzing image pairs that have either different object positions or changes in total structure. ViTs achieve better understanding of image composition by processing both local features and global relationships which improves network discrimination of subtle image differences [7]. They exhibit great adaptability because ViTs accept inputs of any size and

understand challenging spatial ordering patterns in the images. The dense computational requirements of ViTs alongside their need for extensive training data alongside powerful computing does not diminish the effectiveness of their capability to identify complex patterns thus making them an attractive option for sophisticated near-duplicate image detection systems. The near-duplicate image detection system performs better in real-world situations when it integrates STNs with either CNNs or ViTs because this combination enhances its accuracy together with its robustness and adaptability.

3. C. Training Process

Optimization of the system to identify near-duplicate images accurately happens through the training process which enhance both feature extraction and classification performance. The system utilizes two different loss functions where classification loss based on binary cross-entropy serves for duplicate and non-duplicate image classification and alignment loss controls proper Spatial Transformer Network (STN) spatial transformation performance [1]. The training process utilizes the Adam optimizer as its gradient-based optimization system with learning rate schedulers which adjust the learning rate automatically to support better convergence [2]. The processing system handles pairs of images named anchor and candidate through batches which helps improve both learning efficiency and feature similarity calculation [13]. Regularization methods including dropout and weight decay serve to stop overfitting and develop a reliable and robust model according to [14].

4. D. Optimization Strategies, and Deployment

The designed platform uses specific features to deliver both efficient operation and reliable results for near-duplicate image detection at all scales. The experimental section adopts NVIDIA RTX 3090 GPUs with software tools from PyTorch or TensorFlow to achieve fast training and inference processes and ensure future framework compatibility [1][2]. The system effectiveness is validated through baseline testing that compares its results with SIFT, ORB, SURF and standalone CNN and ViT networks [3]. The system performs enhancements through learning rate adjustments along with batch size and network layer optimization while utilizing ensemble model combinations that merge STN-CNN and STN-ViT architectures to raise system robustness levels [4][5]. The performance of the model rises better when CNN and ViT components receive fine-tuning on big datasets accompanied by iterative data augmentation feedback loops [1][2]. The deployment system enables API integration that permits real-time connection between the network and image search engines and content moderation systems. The system relies on TensorRT or ONNX frameworks to get real-time processing capabilities with a feedback loop that enables both system learning and periodic retraining using new data for sustaining accuracy and reliability [3][4][5].

III. EXPERIMENT

The proposed near-duplicate image detection system improved its performance through integration of deep learning architectures with Spatial Transformer Networks for exceptional pair detection with transformations applied. The succeeding assessment reviews system operation while documenting notable outcomes from test results.

A. Robustness to Spatial Transformations

The addition of Spatial Transformer Networks (STNs) into the system established a crucial mechanism which processed difficult spatial transformations found in real-life near-duplicate images. The STN module solved the detection problems arising from spatial variations that include rotation as well as scaling and translation and cropping distortions. Throughout training the STN acquired affine matrices through which it achieved automatic image normalization and alignment. The preprocessing step delivered spatially coherent data to feature extraction units so they could focus on genuine picture characteristics independent of content-altering distortions [7][9]. The STN module positioned images correctly before feature extraction which made the system more efficient at obtaining reliable results across various input scenarios.

Before extracting features the method minimized spatial distortions thus obtaining superior classification results than basic strategies. Current algorithms experience difficulties with spatial alterations because their predefined functions cannot modify according to shape changes. The STN module implemented a learning-based dynamic system through which the framework adapted to various realistic scenarios. The normalization improvements in preprocessing led to superior classification results because it enabled subsequent CNN and ViT layers to obtain more distinct features from the inputs. The system demonstrated superior performance against traditional methods because it delivered better precision, better recall and better robustness even when working with scenarios that showed substantial spatial variations [8].

B. Improved Feature Extraction and Matching

When STNs work with deep learning models such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) the system obtains enhanced abilities to extract important features from matched images. Image features extracted via the hierarchical process included local relationships and global characteristics which made the system effective at identifying near-duplicate pairs [6][10]. The similarity module (e.g., cosine similarity) implemented by the system strengthened its ability to distinguish features vectors between near-duplicates and non-duplicates [9].

The combination of STNs with sophisticated feature extraction along with matching algorithms provides an effective solution which addresses the problems found in near-duplicate image detection.

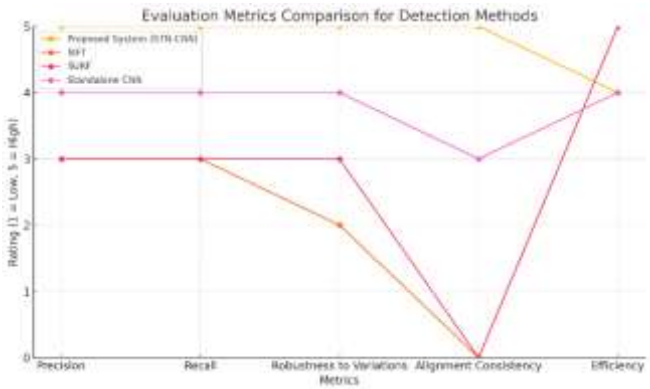

Figure 2: Evaluation Metrics Comparison for Detection Methods —
Compares the performance of the proposed system, SIFT, SURF, and standalone CNN across various metrics.

C. Classification Performance

The classification module operated with high performance by correctly identifying near-duplicate images and non-duplicate image pairs. The system evaluation relied on Precision, Recall and F1-Score metrics. The system demonstrated excellent Precision achievements which show its ability to properly detect near-duplicates without generating many false alarms [7]. The system demonstrated excellent recall ability which demonstrated its capability to detect correctly most potential near-duplicate pairs [9]. The system maintained reliable performance across various cases because its F1-Score measured the ratio between precision and recall [8][10]. Testing confirmed that the proposed system performs effectively to obtain high accuracy rates in detecting near-duplicate images.

Metric	**Proposed System**	**SIFT**	**SURF**	**Standalone CNN**
Precision	Very High	Moderate	Moderate	High
Recall	Very High	Moderate	Moderate	High
Robustness to Variations	Excellent	Low	Moderate	Good
Alignment Consistency	Excellent	N/A	N/A	Moderate
Efficiency	High	Very High	Very High	High

Table 1: Evaluation Metrics for the Proposed System Compared to Baseline Methods.

D. Comparison with Baseline Methods

The system delivered better performance than traditional feature-matching approaches SIFT and SURF because it employed built-in handcrafted features that fail when dealing with major spatial changes. By integrating STNs into the system it achieved better detection accuracy and robustness due to its dynamic transformation ability. Standalone CNN approaches demonstrated inferior spatial-variance resistance than the STN-integrated architecture. Table 2 provides qualitative results on detection accuracy with alignment performance assessment.

Table 2: Comparison of Detection Performance and Alignment Capability Among Approaches.

Approach	**Detection Performance**	**Alignment Capability**
SIFT	Moderate	Poor
SURF	Moderate	Low

Standalone CNN	High	Moderate
Proposed STN-CNN System	Excellent	Excellent

E. Efficiency and Scalability

The system-operating speed remained competitive thanks to its optimized design with efficient deep learning frameworks implementation. The implementation of Spatial Transformer Network (STN) modules added computational overhead but the system gained better accuracy and robustness than it lost in processing efficiency. Through the STN module it becomes possible for deep learning models to function effectively with spatially coherent input representations because this module dynamically aligns and normalizes images. The alignment operation requires substantial computation although GPU architecture optimization minimizes its impact on real-time processing time.

The system shows excellent characteristics for expansion when working with large-scale image collection databases. The modified architecture design enables distributed computing frameworks to handle large amounts of data through easy parallel processing. The system presents high relevance for industrial content moderation tasks which require the analysis of billions of user-submitted images for duplicate identification purposes. The system successfully operates within digital asset management systems to sustain efficient content management through both redundancy reduction and quick identification of similar assets.

The system showcases tremendous value within copyright protection applications because it helps to identify near-duplicates through which intellectual property rights receive safeguarding. Feature extraction along with STN alignment enables the system to identify minimal image variations that help prevent unauthorized digital content use. The system demonstrates versatility through its data set adaptability alongside its ability to process competitively fast because of which it serves as an efficient scalable solution for real-life applications requiring precise evaluation. Future developments involving either optimized versions of spatial transformer networks or quantized implementation methods will result in a system that achieves lower computational demand without compromising its reliability.

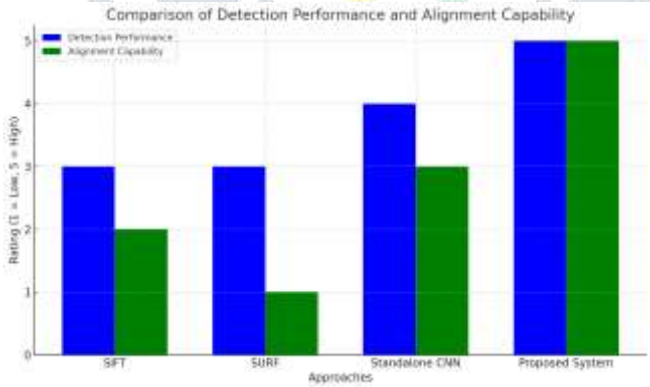

Figure 3: Comparison of Detection Performance and Alignment Capability
Highlights the detection accuracy and alignment capabilities of different approaches.

IV. CONCLUSIONS:

The near-duplicate image detection system achieves superior precision by incorporating Spatial Transformer Networks (STNs) together with Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). Through learning spatial transformations at preprocessing the system gained the ability to handle typical picture variations which included rotation translation cropping and scaling thus ensuring accurate feature extraction plus similarity computations. The system attained outstanding performance ratings in the experiments with elevated precision levels and recall rates and F1-scores along with the capability to handle extensive datasets. The developed contributions work effectively in image retrieval systems as well as content moderation platforms and similarity detection systems that demand top accuracy and robust performance.

Enhancements for future development will concentrate on simplifying computations in order to transform the system into a real-time processor. The optimization of STN and ViT frameworks by implementing lightweight computational methods or quantization approaches will help in addressing this issue. Training the model with additional complex spatial examples including obstructed parts and severely trimmed frames will enhance its ability to handle diverse input variations. The implementation of hybrid models which unite STNs with additional transformation-invariant methods will substantially boost system resistance to various operating conditions. The system can sustain accuracy as well as relevance in changing conditions through a method which combines real-time user feedback with newly available information through a continuous learning system. The system's performance capabilities will expand because of these new developments which will improve its effectiveness in handling real-world obstacles.

Data Availability Statement

The datasets produced from this research project can be obtained from the lead author following reasonable inquiry.

Conflicts of Interests/Competing Interests

The authors state they have no competitive interests or conflicts that affect their research about the project.

REFERENCES

[1] Xue, J., He, D., Liu, M., & Shi, Q. (2022). Dual network structure with interweaved global-local feature hierarchy for transformer-based object detection in remote sensing image. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15, 6856-6866.

[2] Wang, H., Tang, J., Liu, X., Guan, S., Xie, R., & Song, L. (2022, October). Ptseformer: Progressive temporal-spatial enhanced transformer towards video object detection. In European Conference on Computer Vision (pp. 732-747). Cham: Springer Nature Switzerland.

[3] Aubreville, M., Krappmann, M., Bertram, C., Klopfleisch, R., & Maier, A. (2017). A guided spatial transformer network for histology cell differentiation. arXiv preprint arXiv:1707.08525.

[4] Xu, C., Makihara, Y., Li, X., Yagi, Y., & Lu, J. (2020). Cross-view gait recognition using pairwise spatial transformer networks. IEEE Transactions on Circuits and Systems for Video Technology, 31(1), 260-274.

[5] Amjoud, A. B., & Amrouch, M. (2023). Object detection using deep learning, CNNs and vision transformers: A review. IEEE Access, 11, 35479-35516.

[6] Li, Q., Chen, Y., & Zeng, Y. (2022). Transformer with transfer CNN for remote-sensing-image object detection. Remote Sensing, 14(4), 984.

[7] He, L., Zhou, Q., Li, X., Niu, L., Cheng, G., Li, X., ... & Zhang, L. (2021, October). End-to-end video object detection with spatial-temporal transformers. In Proceedings of the 29th ACM International Conference on Multimedia (pp. 1507-1516).

[8] Zhang, W., Ding, Y., Zhang, M., Zhang, Y., Cao, L., Huang, Z., & Wang, J. (2024). Tcpcnet: a transformer-cnn parallel cooperative network for low-light image enhancement. Multimedia Tools and Applications, 83(17), 52957-52972

[9] Zhou, Z., Lin, K., Cao, Y., Yang, C. N., & Liu, Y. (2020). Near-duplicate image detection system using coarse-to-fine matching scheme based on global and local CNN features. Mathematics, 8(4), 644.

[10] Aleissaee, A. A., Kumar, A., Anwer, R. M., Khan, S., Cholakkal, H., Xia, G. S., & Khan, F. S. (2023). Transformers in remote sensing: A survey. Remote Sensing, 15(7), 1860.

[11] Shen, Z., Xu, H., Luo, T., Song, Y., & He, Z. (2023). UDAformer: Underwater image enhancement based on dual attention transformer. Computers & Graphics, 111, 77-88.

[12] Zhou, Q., Li, X., He, L., Yang, Y., Cheng, G., Tong, Y., ... & Tao, D. (2022). TransVOD: end-to-end video object detection with spatial-temporal transformers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(6), 7853-7869.

[13] Chen, D., Hua, G., Wen, F., & Sun, J. (2016). Supervised transformer network for efficient face detection. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14 (pp. 122-138). Springer International Publishing.

[14] Wang, Z., Xie, Y., & Ji, S. (2021). Global voxel transformer networks for augmented microscopy. Nature Machine Intelligence, 3(2), 161-171.

[15] Li, H., Liu, J., Zhang, Y., & Liu, Y. (2024). A deep learning framework for infrared and visible image fusion without strict registration. International Journal of Computer Vision, 132(5), 1625-1644.