Research On Traffic Sign Detection Based on Transformer and Yolov5

Zhonghao Xie

Aberdeen School of Data Science and Artificial Intelligence, South China Normal University, Foshan, 528200, China

3166176865@qq.com

Abstract. YOLOX-Swin algorithm takes Swin-Transformer as the backbone network of YOLOX to extract traffic sign image features, obtain enough global context information through mobile window, and use multiple self-attention mechanism; use YOLOX own path enhanced feature pyramid network to extract and integrate multi-scale feature information including lower information of traffic sign to improve the detection accuracy of small target traffic sign. Because the small target traffic signs occupy fewer pixels in the image, and considering that Transformer requires more training samples than the convolutional network, the original copy and paste method is improved to increase the number of traffic signs samples to further improve the accuracy of object detection. The test results on the TT100K dataset show that the proposed object detection method has higher object detection accuracy than several other methods, and can meet the requirements of accuracy and real-time object detection.

Keywords: Deep learning; YOLOX; Swin-Transformer; small object detection.

1. Introduction

With the continuous development of the computer vision field, object detection is one of the key tasks, aiming to accurately locate and identify objects of interest from images. Early object detection methods often relied on hand-designed features and complex classifiers. However, these methods perform poorly when dealing with complex scenarios and multi-scale targets. Recently, the rise of deep learning techniques has led to remarkable progress in object detection [1]. The proposal of the YOLO algorithm transforms the object detection task into the regression problem, which greatly promotes the development of the object detection field.

However, the traditional YOLO algorithm has some limitations in detecting scenarios such as small and dense targets, which is mainly attributed to its single attention mechanism [2]. To address this problem, researchers began to explore the attention mechanism of introducing Transformer to improve object detection performance [3]. Transformer as a model based on the self-attention mechanism, its successful application in the field of natural language processing indicates its superiority in capturing long-distance dependencies [4]. Recent studies show that applying Transformer to object detection can effectively improve the ability of models to represent targets, thus improving detection accuracy and robustness.

Transformer is a deep neural network based on the self-attention mechanism, which was initially applied in the field of natural language processing [5]. Since the end of 2020, Transformer has been gradually applied in the field of computer vision. Vision in Transformer (ViT) is the first Transformer algorithm applied to image classification [6]. Unlike the invariance and locality of CNN models, the self-attention mechanism of the Transformer model is not limited by local interactions and can learn the inductive bias that best fits the task target [7]. Because the ViT algorithm does not carefully consider the characteristics of visual signals themselves, so it is mainly adapted to image classification tasks, which is not very friendly to region level and pixel level tasks, such as object detection and semantic segmentation [8]. To this end, the academic circle has carried out a lot of improvement work. Among them, the Swin-Transformer algorithm achieves excellent performance in object detection and semantic segmentation tasks.

This paper aims to explore the object detection algorithms based on Transformer and YOLOv5 to compensate for the shortcomings of the traditional YOLO algorithms in dealing with complex scenarios [9]. Specifically, this paper first introduces the core ideas of the YOLOv5 object detection algorithm based on Transformer, including the feature extraction backbone network, the multi-scale feature graph fusion network, and the data enhancement strategy [10]. Subsequently, we performed experiments on the VOC dataset, analyzed the loss function curves, presented the experimental results, and evaluated the performance of the algorithm both quantitatively and qualitatively [11]. Finally, we summarize the experimental results, explore the advantages and disadvantages of the algorithm, and propose possible directions for future improvement.

2. Traffic sign detection and identification algorithm

2.1. Feature extraction backbone network

In object detection tasks, accurate image features with discriminative ability are the basis for efficient detection [12]. The feature extraction backbone network plays a crucial role in this link, and its ability directly affects the model performance [13]. In this paper, the hierarchical Transformer module of Swin-T model can extract the feature maps of different scales. Swin-T's moving window strategy is beneficial to extract the global context information of the image [14]; its multi-head attention mechanism learns the relevant information in their respective representation subspace according to different tasks. In view of this, the backbone network CSPDarknet of the original YOLOX is used to extract multi-scale traffic sign feature information with rich global context information and differentiated features [15]. As shown in Figure 1, Swin-T is located in the backbone network of YOLOv5, Patch partition module receives image information, C2, C3 and C4 groups output feature maps with scales 8080, 4040 and 2020, respectively, and the number of feature channels is 192, 384 and 768, respectively.

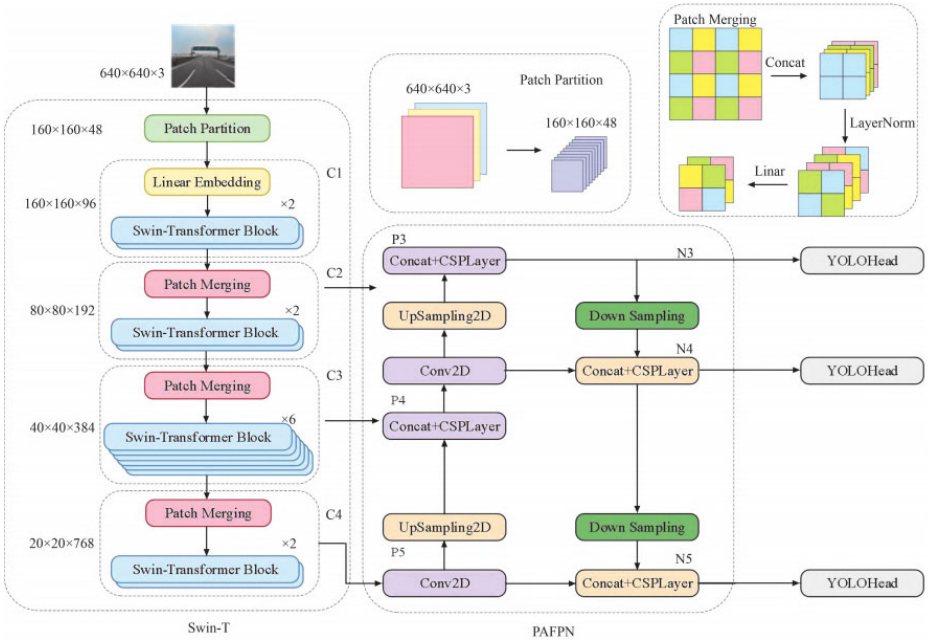

Fig. 1 Network Structure of YOLOv5 based on Swin-T

In the design of feature extraction backbone network, we focus on acquiring multi-scale features of images at different levels. Such multi-scale features have different semantic information and help to identify traffic signs of different sizes and shapes. Therefore, in the upper layer of the network, larger convolution kernels and fewer pooling layers are used to maintain more image detail. In the lower layer, smaller convolution cores and more pooling layers are used to obtain a wider range of receptive fields and semantic information.

2.2. Multi-scale feature graph fusion network

The core idea of multi-scale feature graph fusion network is to integrate feature graphs at different levels, so as to obtain a feature representation with both low-level detail information and high-level semantic information. To achieve this goal, we borrowed the attention mechanism of Transformer and applied it in the process of feature map fusion [16]. The self-attention mechanism of Transformer has the advantage of capturing global information and modeling long-distance dependence and is suitable for feature graph fusion tasks.

Object detection belongs to the small target detection task, which is very sensitive to location information, while the shallow feature map with high resolution can retain more location information and small-scale feature information [17]. As shown in Figure 1, the 8080-feature map of the C2 group carries richer positional information relative to the 2020 and 4040 output by the Score C4 and C3 groups of the backbone network Swin-T.

As shown on the right side of Figure 1, feature maps 2020 and 4040 generate feature integration with the same scale of the same scale with the same scale [18], and P3 and P4. P3 and P4 and P4 and P5 with the same scale through convolution and generate feature maps N4 and N5 after convolution.

PAFPN Network is based on the idea of bidirectional fusion, and constructs top-down, bottom-to-up bidirectional channels [19]. Three feature maps of different scales realize information fusion among different scale features through convolution, upsampling, downsampling and lateral connection. The dimensions of feature maps N3, N4 and N5 are 2020768, 4040384 and 8080192, respectively.

Based on the channel information of the feature plots N3, N4 and N5, and the feature points, the three YOLOHead modules calculate the feature point loss function based on Equation (1) [20], analyze the target coordinates and confidence, and identify the target type. The loss function consists of localization loss L_{reg}, confidence loss L_{obj}, and classification loss L_{cls}. L_{reg} is used to adjust the regression parameters of the prediction box of the feature point; L_{obj} and L_{cls} both use binary cross-entropy loss, L_{obj} is used to adjust whether the prediction box of the feature point contains the target, and L_{cls} is used to adjust the category of the prediction box of the feature point.

$$f_{LOSS} = \frac{L_{cls} + \lambda L_{reg} + L_{obj}}{N_{pos}} \tag{1}$$

Where λ represents the equilibrium coefficient of the localization loss and N_{pos} represents the number of feature points divided into positive samples.

To achieve multi-scale feature map fusion, we introduced position coding in each attention head to retain spatial information [21]. These positional encodes are similar to positional embeddings in Transformer and are able to help the network understand the relationships between different locations in the absence of explicit positional information. By summing position encoding to feature maps, we enhanced the retention of spatial information beyond the attention mechanism.

2.3. Data enhancement strategy

Data augmentation is a commonly used technique in deep learning by transform and expanding the training data to increase the generalization ability and robustness of the model. In the object detection task, the object of data enhancement is to generate new training samples by performing a series of transformations to the input image, so that the model can better adapt to different scenes and perspectives [22]. First, using a random cropping strategy to crop by randomly selecting regions of interest in the image to simulate the position and scale changes of different targets in the image. This helps to improve the position invariance of the model to the target. Secondly, in order to increase the ability of the model to adapt to the scale changes, the multi-scale training (Multi-Scale Training strategy is introduced. During training, the input images were randomly scaled to simulate the size changes of the target at different distances. For the problems existing in traffic sign detection, this

paper adopts the improved copy and paste method to increase the number of small target samples, which includes two steps [23]: first, extract the small target from the original data set, select the effective small target samples and establish a small target sample library; then copy a certain number of small target samples randomly and paste back to the designated area in the original image many times to form a new image, as shown in Figure 2.

In addition, data enhancement strategies such as occlusion and noise are also designed. By introducing occlusion or adding noise in the image, we can simulate the complex situation on the actual road and enhance the robustness of the model for traffic sign identification and positioning.

Fig. 2 Examples of Improved Copy-Paste Algorithm

3. Traffic sign detection experiment

3.1. VOC data set

The VOC dataset series has been released annually since 2005, with versions VOC2005 to VOC2012 etc. These datasets cover 20 common object categories, such as people, vehicles, animals, and furniture, etc [24]. Each category has abundant samples in the dataset and provides detailed annotation information, including object bounding boxes and category labels. Moreover, the VOC dataset also contains challenging scenarios, such as occlusion, attitude change, and background complexity, thus putting higher requirements on the robustness and generalization ability of the algorithm.

VOC datasets have wide implications and applications in object detection research. Its rich annotation information makes it an important benchmark for algorithm performance evaluation, and many advanced object detection algorithms are tested and compared on VOC datasets. The diversity and challenge of VOC datasets have also prompted researchers to propose many innovative methods to tackle different detection problems.

And the VOC dataset plays an important role as one of the standard benchmark datasets to evaluate model performance. Researchers can use the VOC dataset to train and test their models and validate their performance in real scenarios. Moreover, the diversity in the VOC dataset also provides a rich resource for researchers to validate the robustness and generalization ability of the algorithm.

3.2. Loss function curve

The model training loss function curve for the target detection algorithm is shown in Figure 3. Where, "loss_cls" represents the classification loss function curve of the YOLOv5 algorithm on the VOC training set, and "loss_bbox" represents the YOLOv5 algorithm regression loss function curve on the VOC training set.

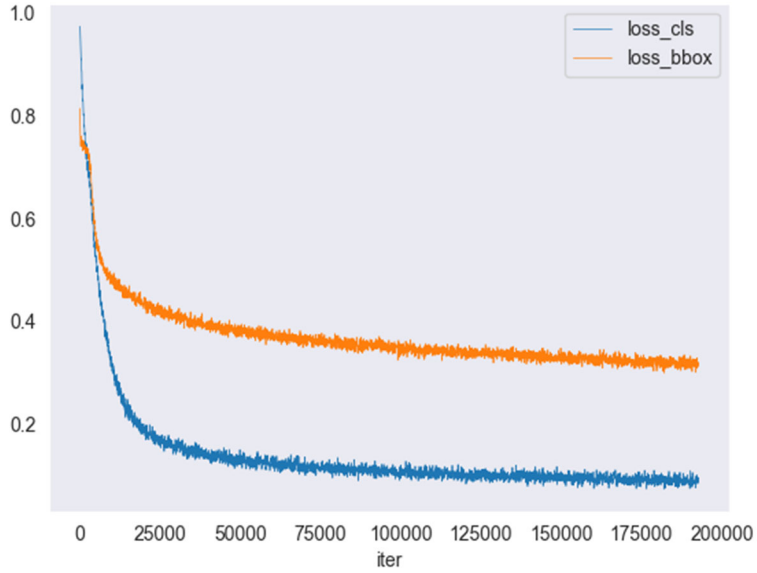

Fig. 3 A loss function curve

Classification loss is a loss function commonly used in object detection tasks to measure the ability of the model to classify different categories [25]. The cross-entropy between its predicted category probability distribution and the true labels was calculated, and then the classification loss of all prediction boxes was summed. In the initial stage of training, the model is weak to classify different categories, so the classification loss is large. As the training proceeds, the model gradually learns the feature representations of different categories, and the classification loss gradually decreases. When the model reaches the convergence state, the classification loss becomes stable, reflecting the ability of the model to accurately classify the various categories.

As Figure 3 shows, the loss_cls curve is at the top and the loss_bbox curve is below the loss_cls curve. The loss_bbox value is always lower than the function value of loss_cls during the training process, indicating that it converges faster. And the loss curve is smoother, and the values fluctuate less.

3.3. Experimental results and their analysis

Experimental results on the VOC dataset show that the Transformer and YOLOv5-based object detection algorithms achieve satisfactory performance in target detection tasks. By testing the VOC dataset, we obtained a range of key performance metrics, such as precision, recall, and F1 scores. Where the precision reflects the correctness of the model in the identified target, the recall rate measures the ability of the model to detect the real target, and the F1 score takes into account the precision and the recall rate comprehensively.

Based on the analysis of the loss function curve, the performance of the model gradually improves with the training. The classification loss and regression loss gradually decrease, indicating that the model gradually learns more accurate target classification and location prediction. This matches the expectations and proves the effectiveness of the adopted Transformer-based YOLOv5 object detection algorithm in the training process.

Table. 1 Test results

Class	Gts	Dets	Recall	Ap
Aeroplane	433	12841	0.984	0.946
Bicycle	358	10388	0.992	0.937
Bird	559	12096	0.996	0.980
Boat	424	20998	0.998	0.905
Bottle	630	31391	0.989	0.917
Bus	301	10663	0.997	0.969
Car	1004	36141	0.994	0.924
Cat	612	8000	1.000	0.991
Chair	1176	43734	0.996	0.887
Cow	298	7670	1.000	0.973
Diningtable	305	15798	0.944	0.799
Dog	759	11221	1.000	0.993
Horse	360	7112	0.997	0.980
Motorbike	356	10149	0.997	0.966
. Person	4372	105315	0.996	0.942
Pottedplant	489	21893	0.988	0.882
Sheep	413	9005	0.995	0.965
Sofa	285	11756	0.989	0.885
Train	315	9194	1.000	0.967
Tvmonitor	392	10702	0.995	0.978
mAP	-	-	-	0.939

3.4. Detection renderings

The output feature layer of the backbone network plays a key role in extracting the features of different scales of the image. We visualize the response situation of these characteristic layers by generating heat maps. On the image, the highlighted area represents the response strength of the feature layer to different targets. These thermal maps will exhibit characteristic responses at different scales, semantics, and locations. At the lower feature layer, we can observe strong responses to local details, while at the higher feature layer the response is more pronounced for higher-level semantic information. This demonstrates the effectiveness of feature extraction backbone networks in extracting multi-level features from images. As shown in Figure 4.

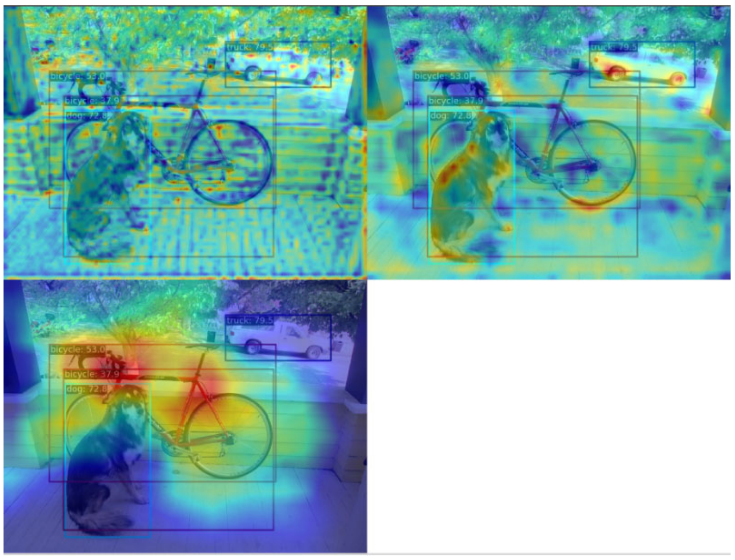

Fig 4 Heat maps of the three output characteristic layers of the backbone network

Multi-scale feature graph fusion networks play a key role in feature fusion. We demonstrate the responses of these characteristic layers through thermal maps to reveal the organic fusion effects of features at different scales. In these thermal maps, the response of different scale features at the corresponding positions is seen, which indicates the success of multiscale feature fusion. Highlighted regions will show the model's ability to locate and classify the landmarks at different scales. As shown in Figure 5.

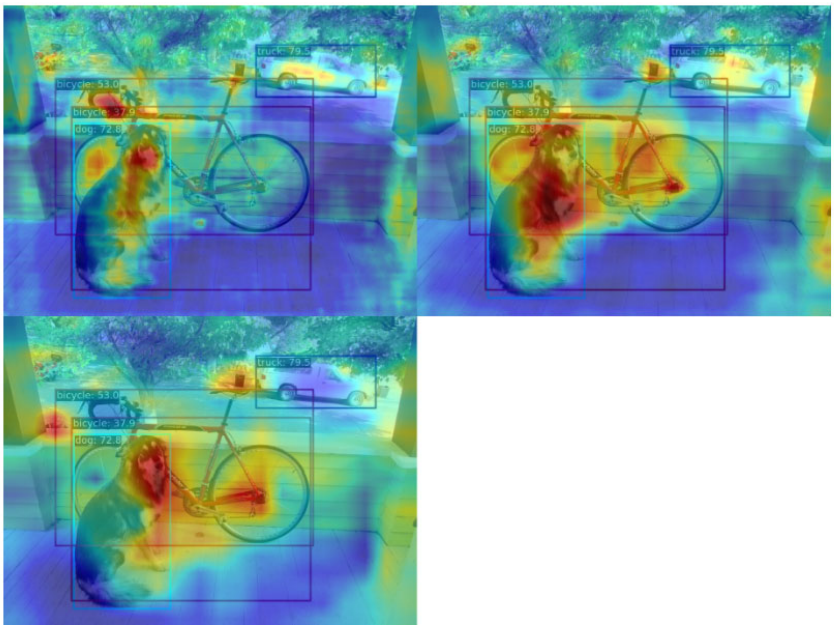

Fig 5 Heat map of the output of the three characteristic layers of novel

4. Summary

For the small target and high-speed accuracy in object detection, a YOLOv5 object recognition algorithm based on double transform is proposed. The backbone network of the lightweight network YOLOv5 is replaced with Swin-T, the multi-scale features are fused with PAFPN, and the training sample is expanded with an improved copy-and-paste method to enhance the performance of Swin-T. Target test experiments on the VOC dataset showed that the YOLOx-Swin algorithm improved by 3.1% on average accuracy and the detection speed by 35% over the Faster R-CNN algorithm. The YOLOX-Swin algorithm improved the average detection accuracy by 3% over the original YOLOX algorithm, 3.2% better than the average detection accuracy of the original YOLOv5 algorithm, and 12.9% better than the average detection accuracy of the original Swin-Transformer algorithm, but the detection speed decreased to 23 frames per second. In conclusion, the YOLOX-Swin algorithm can meet the timeliness and accuracy requirements of traffic sign detection and has a good engineering application prospect.

References

[1] Li Bin, Guo Chenhua. MASPC_Transform: A Plant Point Cloud Segmentation Network Based on Multi-Head Attention Separation and Position Code. Sensors. 2022, 22(23).

[2] Grines Cindy L, Tummala Pradyumna. Will the TRANSFORM II trial transform our management of small vessel coronary disease. Catheterization and cardiovascular interventions: official journal of the Society for Cardiac Angiography. Interventions. 2022,100(5).

[3] Cavalcante Francisco Thálysson Tavares, da Fonseca Aluisio Marques, Holanda Alexandre Jeferson Yves Nunes, dos Santos José C.S. A stepwise docking and molecular dynamics approach for enzymatic biolubricant production using Lipase Eversa® Transform as a biocatalyst. Industrial Crops, Products. 2022, 187(PB).

[4] Yeqing Hu, Guanzheng Tan, Degang Xu, Yaoyi Cai. Grade Prediction of Antimony Concentrate Using Continuous Wavelet Transform and Convolutional Neural Network with Dual Attention. 2022:1104.

[5] Madurapperumage Amod, Johnson Nathan, Thavarajah Pushparajah, Tang Leung, Thavarajah Dil. Fourier Ttransform infrared spectroscopy (FTIR) as a high throughput phenotyping tool for quantifying protein quality in pulse crops. The Plant Phenome Journal. 2022, 5(1).

[6] Galland J. Artificial intelligence will transform internal medicine. La Revue de medecine intern. 2022, 43(5).

[7] Kraisiri KHIDKHAN, Amnart POAPOLATHEP, Saranya POAPOLATHEP, Sittinee KULPRASERTSRI. Residues of neonicotinoids and fipronil in paddy fields and duck eggs: Do ducks transform and accumulate these substances to egg products? 2022, 49.1(0).

[8] Carvalho Wagner C.A., Luiz Jaine H.H., Fernandez-Lafuente Roberto, Hirata Daniela B., Mendes Adriano A. Eco-friendly production of trimethylolpropane triesters from refined and used soybean cooking oils using an immobilized low-cost lipase (Eversa>® Transform 2.0) as heterogeneous catalyst. Biomass and Bioenergy. 2021, 155.

[9] Sun Shangde, Guo Jingjing, Chen Xiaowei. Biodiesel preparation from Semen Abutili (Abutilon theophrasti Medic.) seed oil using low-cost liquid lipase Eversa® transform 2.0 as a catalyst. Industrial Crops; Products. 2021, 169.

[10] Anna Kostianko, Sergey Zelik. Kwak Transform and Inertial Manifolds revisited. Journal of Dynamics and Differential Equations. 2021, 34(4).

[11] Pradeep S, Nirmaladevi P. A. Review on Speckle Noise Reduction Techniques in Ultrasound Medical images based on Spatial Domain, Transform Domain and CNN Methods. IOP Conference Series: Materials Science and Engineering. 2021, 1055(1).

[12] Aker. BP and Cognite Update on Robotics Deployment Offshore to Transform Oil and Gas Industry with Autonomous Mission. Manufacturing Close - Up, 2020.

[13] Bogdan Ciprian. Book Review: Noile subiectivități ale capitalismului global. Spiritualitate, dezvoltare personală și transformări neoliberale în România (The New Subjectivities of Global capitalism. Spirituality, Personal Development and Neoliberal Transformations in Romania), Sorin Gog and Anca Simionca (eds.), Cluj-Napoca: Tact. Studia Universitatis Babes-Bolyai Sociologia. 2020, 65(2).

[14] Health and Medicine - Medical Physics; Studies from University of Massachusetts in the Area of Medical Physics Reported (Locally Linear Transform Based Three-dimensional Gradientl0-norm Minimization for Spectral Ct Reconstruction). Journal of Mathematics. 2020.

[15] Aruba a Hewlett Packard Enterprise company; Montreal's Saputo Stadium Chooses Aruba to Transform the Gameday Experience for MLS Fans. Telecommunications Weekly. 2020.

[16] North Tees and Hartlepool NHS Foundation Trust chooses Navenio's location tech to transform service lines. M2 Presswire. 2020.

[17] Mathematics: Researchers at Indian Institute of Technology (IIT) Gandhinagar Release New Data on Mathematics [Generalized Lambert Series, Raabe's Cosine Transform and a Generalization of Ramanujan's Formula for Zeta (2m+1)] [J]. Journal of Mathematics,2020.

[18] Tang Haozhe,Du Shuchun,Zeng Qingqin,Li Shuo,Hu Fei. Study and application on coherent noise suppression by curvelet transform. 2020:467-469.

[19] I don't understand why Africa is still hungry: UN envoy's plan to transform food systems for all. 2020, M2 Presswire.

[20] North Carolina State University: Landmarks facing climate threats could 'transform,' expert says NewsRx Health. Science,2020.

[21] Phillips 66; Phillips 66 Plans to Transform San Francisco Refinery into World's Largest Renewable Fuels Plant. Ecology Environment; Conservation. 2020.

[22] QUALCOMM Incorporated; Patent Application Titled "Secondary Transform Designs for Partitioned Transform Units in Video Coding" Published Online (USPTO 20200252622). Computer Weekly News. 2020.

[23] Physics Astrophysics: Reports from Canada-France-Hawaii Telescope Corporation Add New Data to Findings in Astrophysics (Wide Field-of-view Study of the Eagle Nebula with the Fourier Transform Imaging Spectrograph Sitelle at Cfht). Journal of Physics Research. 2020.

[24] DXC Technology; Sabre Selects DXC Technology to Help Transform the Future of Travel as Part of a Multi-Year Renewal of the Companies' Agreement. Technology News Focus. 2020.

[25] Engineering: Studies from Xi'an Polytechnic University Add New Findings in the Area of Engineering (Noise Reduction on Received Signals in Wireless Ultraviolet Communications Using Wavelet Transform). Journal of Engineering. 2020.