




GridFormer: Grid Foreign Object Detection also Requires Transformer

Dengquan Wu¹(✉) , Hanxin Zheng¹, Zixi Li¹, Xin Xie¹, Tijian Cai¹, and Fengyang Shang²

¹ East China Jiaotong University, Nanchang 330000, China
2310555878@qq.com

² Columbia University, New York, USA

Abstract. Under the rapid development of China's energy industry, the invasion of foreign objects poses a considerable challenge to the operation and maintenance of power transmission channels. Due to the particularity of the power system, there needs to be more open-source datasets for power transmission foreign object detection, which limits the development of this field and needs to be addressed urgently. Additionally, existing object detection models are often too complex to meet the real-time inference requirements of drones and other terminal devices. To solve these problems, this paper proposes a lightweight object detection model named GridFormer based on a hybrid feature extraction network. This model combines the advantages of convolutional neural networks (CNNs) and transformers, aiming to improve object detection accuracy and real-time performance. Experimental results demonstrate that the proposed model achieves an mAP value 96.78 on a power transmission foreign object dataset. On an NVIDIA GPU 3080, the inference speed can reach 68.7 FPS with only minor loss compared to GhostNet. The model achieves an mAP value of 79.04 on the Pascal VOC dataset, further validating its effectiveness. Compared to GhostNet, the proposed model exhibits superior performance in terms of object detection performance. By addressing the issues above, the GridFormer model is expected to support the development of China's energy industry, improve the efficiency of power transmission operation and maintenance, and promote the development of the object detection field.

Keywords: foreign object detection · Transformer · lightweighting · hybrid features · Grid system

1 Introduction

Electricity, as a crucial essential industry, significantly impacts people's lives and property safety, the national economy, and overall economic development

Supported by the National Natural Science Foundation of China, under Grant No. 62162026, and the Science and Technology Project supported by the Education Department of Jiangxi Province, under Grant No. GJJ210611 and the Science and Technology Key Research and Development Program of Jiangxi Province, under Grant No. 20203BBE53029.

[1]. However, due to long-term exposure to natural elements such as rain, snow, and foreign substances, power transmission lines are prone to issues such as foreign object attachments [2]. If these issues are not addressed promptly, they can compromise the stability of the power grid. Current solutions often involve computer vision techniques to detect foreign objects on power transmission lines. For example, Zhang proposed a method named RCNN4SPTL, which replaces the feature extraction network of Faster RCNN with a more lightweight SPTL-Net, thereby improving the detection speed [3]. Additionally, Wang et al. proposed a method based on SSD for power transmission line detection, studying the effects of different feature extraction networks and network parameters on the accuracy and speed of object detection [4]. Jiang et al. identified bird nests on power transmission lines, cropping the detection results into sub-images and filtering out those not containing bird nests using an HSV color space model, significantly improving detection accuracy [5]. Qiu focused on birds as an object category and proposed a lightweight YOLOv4-tiny network model for bird detection on power transmission lines, providing a basis for preventing bird-related power grid shutdowns [6]. However, in real-world scenarios, foreign objects that can invade power transmission lines include birds, bird nests, balloons, and kites [7]. The high-resolution images captured by drones and other equipment contrast with the relatively small size of these foreign objects, posing a challenge for foreign object intrusion detection. This article proposes a lightweight hybrid object detection network named GridFormer that combines convolution’s inductive bias advantages and transformers’ global modeling capabilities [8] to achieve good generalization performance on object detection tasks [9, 10]. This method strikes a new tradeoff between computational cost and detection accuracy, providing a new, effective solution for foreign object intrusion detection in power transmission lines.

2 Methods

2.1 Model Design

This paper intends to construct a lightweight object detection model based on the CNN-Transformer network, see Fig. 1. It consists of three parts. The first part is a hybrid feature extraction network, which extracts image semantic information; the second part is a feature fusion part, which fuses different levels of feature maps through up-sampling and down-sampling operations to generate a multi-scale feature pyramid. The last part is the classification and localization part, which introduces the auxiliary head design and dynamic label assignment strategy so that the middle layer of the network learns more information by richer gradient information to help train. This model combines the advantages of convolution and transformer to seek Pareto improvement in computational cost and detection accuracy on the object detection task.

In this paper, we use the self-attention mechanism in the deep part of the network, which avoids the Patch division of the feature map due to its small enough

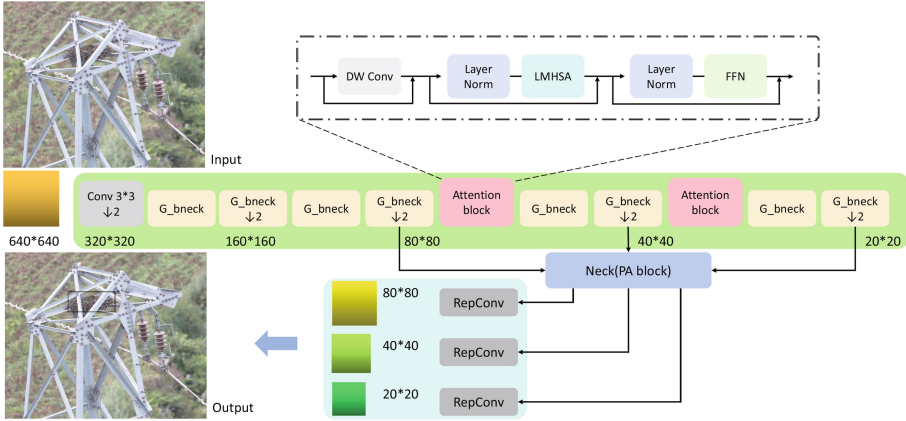


Fig. 1. The overall framework of GridFormer.

size, reduces the effect of positional coding, and ensures the overall consistency of the feature map.

2.2 Attention Block

The structure of the Attention Block is modeled after the design of the Transformer encoder, which is shown in the dashed box in Fig. 1. Absolute position encoding is usually used in ViT, and each patch corresponds to a unique position encoding, so it is not possible to achieve translation invariance in the network. Attention Block uses a 3×3 depth-separated convolution introduces translation invariance of the Transformer module, and stabilizes the network training using residual connectivity. Furthermore, the computational procedure is shown in Eq. 1:

$$f(x) = DWConv(x) + x \tag{1}$$

The vital design in the Attention Block is LMHSA. Given an input of size $\mathbb{R}^{n \times c}$, the original multi-head attention mechanism first generates the corresponding Query, Key, and Value. Then, by dot-producing the point Query and Key, it produces a weight matrix of size $\mathbb{R}^{n \times n}$:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right) * V \tag{2}$$

This process tends to consume a lot of computational resources due to the large size of the input features, making it difficult to train and deploy the network. We use kxk average pooling kernel to downsample the Key and Value branch generation. The calculation process of the lightweight self-attention mechanism is shown in Fig. 2:

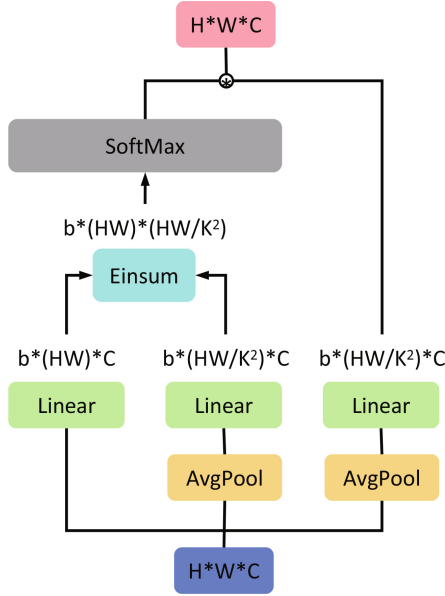


Fig. 2. Calculation process diagram of lightweight self-attention mechanism.

Two relatively small feature maps K' and V' were obtained.

$$K' = AvgPool(K) \in \mathbb{R}^{\frac{n*c}{k^2}} \quad (3)$$

$$V' = AvgPool(V) \in \mathbb{R}^{\frac{n*c}{k^2}} \quad (4)$$

where k is the size and sampling step of the pooling kernel, n is the product of input feature maps H and W , and c is the number of feature map channels. Introducing a pooling layer in the Self-Attention module to downsample the feature map efficiently saves computation and memory. Moreover, LMHSA is composed of multiple Lightweight Self-Attention, see Eq. 5:

$$LMHSA(x) = Concat_{i=[1:h]}[softmax(\frac{Q_i K_i^T}{\sqrt{d}}) * V_i] \in \mathbb{R}^{n*d} \quad (5)$$

The computational complexity of LMHSA is:

$$O(LMHSA) = h * (\frac{n * c^2}{k^2} + \frac{n^2 * c}{k^4} + O_\phi(\frac{n^2}{k^2})) \quad (6)$$

where n is the product of the length and width of the output feature map, h is the number of heads of the multi-head self-attention mechanism, k is the kernel size and step size of the pooling kernel, c is the number of channels of the input feature map and the output feature map, and ϕ denotes the softmax activation function.

The computational complexity of the standard self-attention mechanism can be expressed as:

$$O(LMHSA) = h * (n * c^2 + n^2 * c + O_\phi(n^2)) \quad (7)$$

Compared with MHSA, the computational cost of LMHSA is about $\frac{1}{K^2}$ of that of MHSA. The LMHSA in this paper effectively reduces the computational cost, and the optimized matrix operation is more friendly to network model training and inference.

3 Training Methods

In order to improve the accuracy and generalization ability, this paper introduces mixup [13], Mosaic [14] data augmentation, cosine annealing [15] and label smoothing [16] to train the model.

3.1 Data Augmentation

The currently available foreign object dataset has limited capacity, and the size of the sample capacity is crucial for the training effect of the network model. Therefore, we can use data augmentation methods to expand the dataset. Among the commonly used methods for data augmentation include techniques such as spatial transformation and colour transformation. In addition, this paper uses data augmentation methods such as Mosaic and Mixup of image mixing classes. Mosaic data augmentation improves the CutMix method, which aims to enrich the image background while improving the model’s detection performance for small objects.

3.2 Cosine Annealing

In order to avoid the model from falling into local optimal solutions, a learning rate adjustment strategy can be used, and one of the standard methods is cosine annealing. The learning rate is gradually reduced through cosine annealing so that the model can better search for the global optimal solution during the training process. This strategy is widely used in deep neural network training and has achieved good results. The learning rate can be reduced by the cosine annealing function, denoted as

$$\eta_t = \eta_{min}^i + \frac{1}{2}(\eta_{max}^i - \eta_{min}^i)[1 + \cos(\frac{T_{cur}}{T_i}\pi)] \quad (8)$$

where η_{max} and η_{min} are the maximum and minimum values of the learning rate, respectively, and T_{cur} and T_i are the current and total number of iterations of an epoch, respectively.

3.3 Label Smoothing

One-hot coded labels in multiclassification problems tend to lead to model overfitting because the model focuses on probability values close to 1. To address this problem, label smoothing can be used to balance the model’s predictions and reduce the risk of overfitting. Label smoothing is introduced to smooth the categorical labels, denoted as:

$$y'_i = y_i(1 - \varepsilon) + \frac{\varepsilon}{M} \quad (9)$$

where y'_i is the label after label smoothing, y_i is the one-hot label encoding, M is the number of categories, and ε is the label smoothing hyperparameter.

3.4 Anchor Clustering

The scheme chosen in this paper is an anchor-based object detection model. The traditional anchor selection method is challenging to improve the accuracy, and we first use the K-means clustering algorithm to cluster the manually labelled actual bounding boxes in the training set to obtain the optimal anchor size. Then, we select nine anchor points to predict the bounding box based on the average IoU to improve the detection accuracy. The clustering steps are:

1. Randomly select N boxes as initial anchors;
2. Using the IOU metric, assign each box to the ANCHOR that is closest to it;
3. Calculate the mean value of the width and height of all boxes in each cluster and update the position of the anchor;
4. Repeat steps 2 and 3 until the anchor no longer changes or the maximum number of iterations is reached.

The anchor clustering centers of the transmission line foreign object intrusion dataset in this paper were calculated by K-means as [[38, 49], [81, 55], [61, 88], [78, 153], [110, 120], [144, 98], [155, 182], [200, 257], [317, 380]], as illustrated in Fig. 3. Figure 4 illustrates the Pascal VOC dataset category of anchor clustering.

4 Experiments

4.1 Datasets

Significantly, few open-source datasets are related to the grid intrusion of foreign objects. The foreign object detection dataset used in this paper is mainly from the dataset provided by the 2nd Guangzhou-Pazhou Algorithm Competition-Complex Scene-Based Transmission Corridor Hidden Dangerous object Detection Competition [11], with a total of 800 annotated image data, and the ratio of this paper’s training set and test set division is 9:1. among the categories of foreign objects are nest, balloon, kite and trash. This paper also validates the effectiveness of this paper’s model on the open-source dataset Pascal VOC [12]. Pascal VOC 2007 has 9,963 images containing 24,640 labelled objects, and Pascal VOC 2012 has 11,540 images containing 27,450 labelled objects, which contain the same 20 object classes for the object detection task.

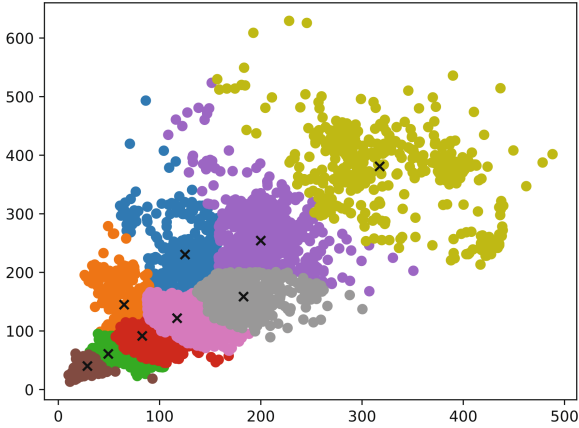


Fig. 3. Anchor clustering of transmission line foreign object detection dataset.

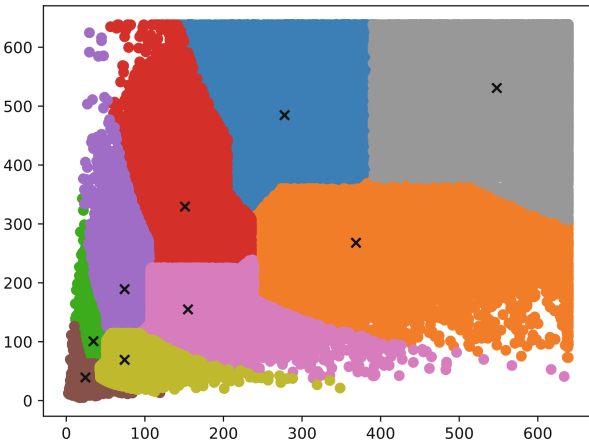


Fig. 4. Anchor clustering of Pascal VOC dataset.

4.2 Experimental Settings

In this paper, the GridFormer model is constructed based on the PyTorch framework, and the neural network parameters are optimized using the Adam optimizer. The initial learning rate is $1e-3$, the minimum learning rate is $1e-5$, cosine annealing is used to attenuate the learning rate, label smoothing is set to 0.005, the input resolution is $640 * 640$, the batch size is set to 8, and the maximum epoch is set to 100. All the training is done using NVIDIA RTX 3080 GPU.

4.3 Evaluation Metrics

In order to reasonably evaluate the performance of the lightweight object detection model, this paper adopts the average value of the APs of each category (mAP) to measure the detection accuracy of the object detection model; the number of floating-point operations, GFLOPs, is used to measure the computational amount of the model, and the number of parameters, Params (M), is used to measure the complexity of the model, which together reflect the computational cost of the model. The number of frames per second, FPS, is used to measure the inference speed of the model. The above four metrics are used to determine the trained model’s performance comprehensively.

4.4 Results

On the target dataset of transmission channel hazards, this paper has done relevant experiments and completed the test; the experimental results are shown in Table 1. GridFormer achieved 96.78% mAP with only 25.39M parametric quantities, and the P-R curves of the four categories in the dataset are shown in Fig. 5.

In this paper, we compare the excellent network design of Ghostnet. GridFormer improves the mAP by 4.96% over Ghostnet in terms of accuracy, and the model can reach 68.7 FPS in terms of inference speed, which can satisfy the demand for real-time transmission line foreign object detection.

Table 1. Performance of GridFormer on transmission line foreign object detection dataset

Model	GFLOPs	Params(M)	mAP(%)	FPS
Ghostnet	20.86	26.75	91.82	75.1
GridFormer	21.16	25.39	96.78	68.7

This paper also validates the effectiveness and generalization performance of GridFormer on the Pascal VOC dataset, and the experimental results are shown in Table 2.

Compared with Ghostnet, GridFormer improves the AP on all 15 categories, in which the accuracy of the cow category improves more than 10 AP, and the detection accuracy exceeds 91 AP for both the aeroplane and horse categories. In this paper, we found that the four categories with lower AP of bottle, chair, diningtable, and pottedplant have the superclass of Household, which is due to the cumbersome categories, severe occlusion in household scenarios, and the significant personalized differences of the detected categories, etc. leading to the poor performance in the model inference.

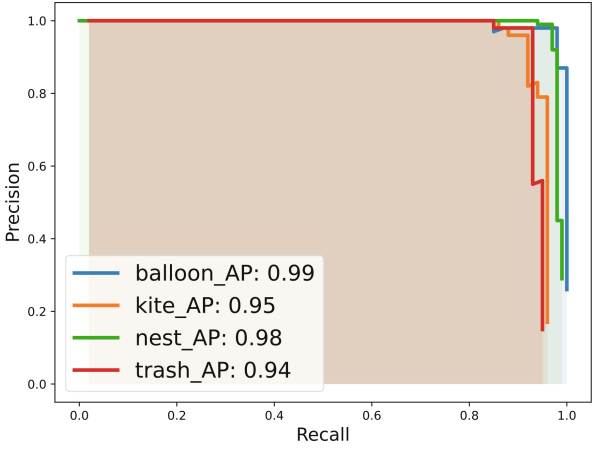


Fig. 5. P-R curves for each category of transmission line foreign object detection dataset.

Table 2. Performance of GridFormer on Pascal VOC dataset

Category	Number of labels	Ghostnet	GridFormer
aeroplane	1336	89.29	91.17
bicycle	1259	87.53	88.64
bird	1900	82.71	82.58
boat	1294	71.94	68.95
bottle	2013	57.62	59.37
bus	911	88.56	87.43
car	4025	83.29	86.20
cat	1768	85.94	83.05
chair	3497	61.53	61.99
cow	1012	70.77	81.89
diningtable	922	61.98	63.01
dog	2222	81.75	81.15
horse	1287	88.99	91.14
motorbike	1248	84.00	87.95
person	16051	88.31	88.86
pottedplant	1784	53.14	53.18
sheep	1129	84.29	83.63
sofa	951	71.84	73.01
train	1086	82.44	82.99
tvmonitor	1283	81.28	84.56
mAP	/	77.86	79.04

4.5 Ablation Studies

This paper experimentally compares the effect of different training methods on the results. The effects of data augmentation, cosine annealing, and label smoothing on the object detection results are further explored.

Table 3 demonstrates the effect of training methods on model accuracy and F1 scores. In Experiments 2, 3, and 4, Mosaic and Mixup data augmentation techniques were introduced to be able to generate new samples, and the comparison between Experiment 1 and Experiment 2 without the addition of such data augmentation methods showed a growth of 1.4% mAP compared to Experiment 1 without such data augmentation methods. The comparison between Experiment 2 and Experiment 3 shows that labelling improves the model accuracy by 0.6% mAP. In Experiment 4, the label smoothing technique is introduced in this paper, which improves the model accuracy by 1.97% mAP. We believe that this is related to the existence of a more severe data imbalance in the dataset, where there is more data in the category of nest, and the label smoothing technique better handles the scarcity of the samples by adjusting the probability distributions of category labels, thus improving the model’s accuracy for all categories. The situation, thus improving the model’s ability to detect all categories.

Table 3. Effect of data augmentation, cosine annealing, and label smoothing on the detection.

Exp	method	mAP(%)	F1-Score
1	none	92.80	0.893
2	DA	94.20	0.935
3	DA+CA	94.81	0.933
4	DA+CA+LS	96.78	0.955

DA: data augmentation; CA: cosine annealing; LS: label smoothing.

5 Conclusion

In grid transmission line systems, foreign object detection is an important protection measure to ensure the regular operation of transmission lines. In this paper, we propose a lightweight object detection model GridFormer based on a CNN-transformer hybrid feature extraction network, which effectively combines the advantages of convolutional induction bias and the transformer’s long-term dependence and still can show the excellent performance of the transformer on smaller datasets. The application of GridFormer in the field of transmission line foreign object detection can be better adapted to the scenario of diverse foreign object morphology, smaller objects to be detected and smaller data volume. The experiments show that the model in this paper can find a new tradeoff between inference speed and detection accuracy.

Acknowledgment. This paper is supported by the National Natural Science Foundation of China, under Grant No. 62162026, and the Science and Technology Project supported by the Education Department of Jiangxi Province, under Grant No. GJJ210611 and the Science and Technology Key Research and Development Program of Jiangxi Province, under Grant No. 20203BBE53029.

References

1. Notice of the National Energy Administration on Issuing the “14th Five Year Plan for Electric Power Safety Production”. http://zfxgk.nea.gov.cn/2021-12/08/c_1310442211.htm
2. Ai, Z.: Research on Anomaly Detection Algorithm of Overhead Transmission Lines Based on UAV Aerial Images, Northeast Electric Power University (2021)
3. Zhang, W., Liu, X., Yuan, J.: RCNN-based foreign object detection for securing power transmission lines (RCNN4SPTL). *Procedia Comput. Sci.* **147**(1), 331–337 (2019)
4. Liang, H., Zuo, C., Wei, W.: Detection and evaluation method of transmission line defects based on deep learning. *IEEE Access* **8**, 38448–38458 (2020)
5. Hao, J., Wulin, H., Jing, C., Xinyu, L., Xiren, M., Shengbin, Z.: Detection of Bird Nests on Power Line Patrol Using Single Shot Detector, Chinese Automation Congress (CAC), Hangzhou, China, pp. 3409–3414 (2019)
6. Qiu, Z.B., Zhu, X., Liao, C.B.: Detection of bird species related to transmission line faults based on lightweight convolutional neural network. *IET Gener. Transm. Distrib.* **16**(1), 869–881 (2022)
7. Liu, J.: Research on foreign object detection algorithm and software design of transmission line based on YOLOX, China University of Mining and Technology (2023)
8. Vaswani, A., Shazeer, N., Parmar, N.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000–6010 (2017)
9. Chen, Y.P., Dai, X.Y., Chen, D.D.: Mobile-former: bridging MobileNet and transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5270–5279 (2022)
10. Gulati, A., Qin, J., Chiu, C.-C.: Conformer: convolution-augmented transformer for speech recognition. [arXiv:1706.03762](https://arxiv.org/abs/1706.03762) (2021)
11. Transmission Channel Hidden Dangerous Object Detection Algorithm Based on Complex Scenarios. <https://aistudio.baidu.com/competition/detail/952/0/introduction>
12. Everingham, M., Gool, L.V., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.* **88**(2), 303–338 (2010)
13. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: beyond empirical risk minimization. In: ICLR 2018 Conference Blind Submission (2018)
14. Bochkovskiy, A., Wang, C.-Y., Liao, H.: YOLOv4: optimal speed and accuracy of object detection. [arXiv:2004.10934](https://arxiv.org/abs/2004.10934) (2020)
15. Xu, G., Cao, H., Dong, Y., Yue, C., Zou, Y.: Stochastic gradient descent with step cosine warm restarts for pathological lymph node image classification via PET/CT images. In: 2020 IEEE 5th International Conference on Signal and Image Processing (ICSIP), Nanjing, China, pp. 490–493 (2020)
16. Wang, C.-Y., Bochkovskiy, A., Liao, H.-Y.M.: YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. [arXiv:2207.02696](https://arxiv.org/abs/2207.02696) (2022)