Explore Innovative Depth Vision Models with Domain Adaptation

Wenchao Xu[1], Yangxu Wang[2]
School of Electrical and Computer Engineering, Nanfang College Guangzhou, Conghua 510970, China[1]
Department of Network Technology, Software Engineering Institute of Guangzhou, Conghua 510990, China[2]

***Abstract*—In recent years, deep learning has garnered widespread attention in graph-structured data. Nevertheless, due to the high cost of collecting labeled graph data, domain adaptation becomes particularly crucial in supervised graph learning tasks. The performance of existing methods may degrade when there are disparities between training and testing data, especially in challenging scenarios such as remote sensing image analysis. In this study, an approach to achieving high-quality domain adaptation without explicit adaptation was explored. The proposed Efficient Lightweight Aggregation Network (ELANet) model addresses domain adaptation challenges in graph-structured data by employing an efficient lightweight architecture and regularization techniques. Through experiments on real datasets, ELANet demonstrated robust domain adaptability and generality, performing exceptionally well in cross-domain settings of remote sensing images. Furthermore, the research indicates that regularization techniques play a crucial role in mitigating the model's sensitivity to domain differences, especially when incorporating a module that adjusts feature weights in response to redefined features. Moreover, the study finds that under the same training and validation set configurations, the model achieves better training outcomes with appropriate data transformation strategies. The achievements of this research extend not only to the agricultural domain but also show promising results in various object detection scenarios, contributing to the advancement of domain adaptation research.**

Keywords*—Deep learning; neural network; domain adaptation; lightweight; regularization techniques

I. Introduction

Remote sensing technology, as a crucial tool for observing the Earth and its ecosystems, has played an indispensable role in multiple domains [1]. However, due to various factors such as capture devices, time, and location influencing the acquisition process of remote sensing images, there exist differences in remote sensing data across different domains. Consequently, models trained in one domain (source domain) often exhibit decreased performance when applied to another domain (target domain). This challenge is commonly referred to as distributional difference [2].

Indeed, distributional difference has long been a persistent issue in machine learning. A series of studies have demonstrated that as the mismatch between distributions increases, performance noticeably declines [3]. A widely adopted approach to address this issue is Domain Adaptation (DA). Previous research has shown that domain adaptation markedly impacts the accuracy and reliability of processing remote sensing images. To tackle this problem, researchers have explored various methods, such as reannotating a portion of target domain data for model fine-tuning [2], making the data distributions of the source and target domains more similar through feature selection or transformation [4], utilizing adversarial training for domain alignment [5], and employing strategies like self-supervised learning [6] and meta-learning [7]. These approaches have, to some extent, alleviated the problem of distributional differences.

With the rapid development of the third wave of artificial intelligence—deep learning, deep convolutional neural networks (CNNs) are significantly pushing the performance boundaries of computer vision at an incredible pace [8]. The latest advances in Unsupervised Domain Adaptation (UDA) in image processing have been attempted and progressed in various fields. Goel et al. [9] achieved unsupervised domain adaptation by guiding transfer learning and employing the Jensen-Shannon (JS) divergence method. In the remote sensing domain, Elshamli et al. [10] introduced an innovative approach to domain adaptation, incorporating denoising autoencoders and domain adversarial neural networks, especially in the classification of hyperspectral and multispectral images. In the agricultural domain, Zhang et al. [11] narrowed the gap between the source and target domains for agricultural land extraction using Generative Adversarial Networks (GANs). Similarly, Valerio et al. [12] accomplished unsupervised leaf counting, while Marino et al. [13] achieved potato defect classification. In plant disease recognition, Fuentes et al. [14] proposed open-set adaptation and cross-domain adaptation methods to enhance tomato disease recognition using unlabeled data. Additionally, Wu et al. [15] achieved cross-domain recognition of wild plant diseases. In robotics, Magistri et al. [16] introduced Unsupervised Domain Adaptation (UDA) techniques for semantic segmentation, enhancing the adaptability of agricultural robots to better perceive and understand different environments. These collective efforts address challenges associated with domain transfer and variability, contributing to the robustness and adaptability of image processing models for various application domains.

Despite these advancements, it is noteworthy that the visual representations learned by deep CNNs exhibit considerable domain invariance. There is evidence that combining existing CNN representations with a linear classifier can achieve relatively high accuracy [17]. In earlier research, Lu et al. [18] posed a challenging question, namely achieving domain adaptation without explicit adaptation of data distribution. This

provided an opportunity to reconsider the problem of cross-domain generalization from a new perspective.

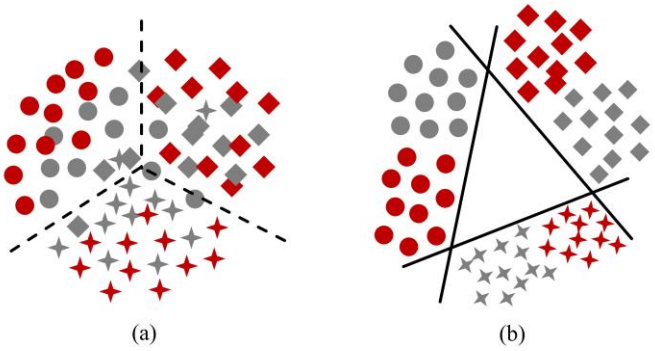

Fig. 1. Two typical scenarios. The target domain is shown in gray and the source domain is shown in red. Different tags indicate different categories.

As Fig. 1 illustrates two typical data distributions, an interesting phenomenon can be observed: samples from different domains but of the same class are close enough, while samples from different classes have a sufficiently large gap. In scenario (a), the distribution difference between the source and target domains is small, but due to confusion among different classes, the final recognition performance is not ideal. In contrast, in scenario (b), despite a marked difference between different domains, samples from different classes are still linearly separable. This suggests that the magnitude of differences between the source and target domains is not a good indicator of the final recognition accuracy. Is the CNN representation powerful enough to eliminate the need for domain adaptation? This study aim to explore a visual model that achieves domain adaptation without the need for explicit adaptation, designed a novel deep convolutional neural network, Efficient Lightweight Aggregation Network (ELANet), which innovatively employs an ELA module for optimizing feature decoding. Experimental results demonstrate the significant optimization of ELANet, showcasing domain adaptability due to the powerful representational capabilities of current CNNs and some carefully designed features. The validation process is based on three datasets: the Global Wheat Head Detection 2021 (GWHD) dataset focusing on wheat [19], and two remote sensing datasets, namely Remote Sensing Object Detection (RSOD) [20] and University of Chinese Academy of Sciences - Aerial Object Detection (UCAS-AOD) [21]. For the remote sensing datasets, the "airplane" category was selected, with RSOD serving as the source domain and UCAS-AOD as the target domain. Extensive experimental results demonstrate the effectiveness of the ELANet method, and some interesting findings are reported.

In summary, the contributions of the research can be summarized as follows:

- ELANet: A visual model with domain adaptation, reporting state-of-the-art performance in cross-domain settings for agricultural and remote sensing scenarios, demonstrating sufficient generality.
- Validation of the effectiveness of regularization techniques: The study proves that utilizing regularization techniques to mitigate domain variance is effective. In particular, incorporating a module that dynamically adjusts feature weights in response to domain redefinition is a more intelligent approach.
- Effective Training Set and Validation Set Configuration: Training under the same configuration for training and validation sets is more effective, provided the existence of data transformation strategies.

II. Methodology

A. ELANet Model Design

The ELANet model is designed based on two components: Encoder and Decoder. The Encoder extracts and downsamples features from input images through a series of convolutional and channel transformation layers, forming feature maps at different resolutions. The Decoder is responsible for the feature extraction task and includes multiple branches. Each branch processes features at different scales through convolutional and channel fusion operations. Multi-scale feature fusion enhances detection performance. Finally, an Adaptive Scale Fusion (ASF) layer is employed to adaptively fuse features, reducing the need for deep downsampling to obtain high-level semantic information [22], as depicted in Fig. 2. The following sections will introduce the global architecture of the ELANet model and its optimizations.

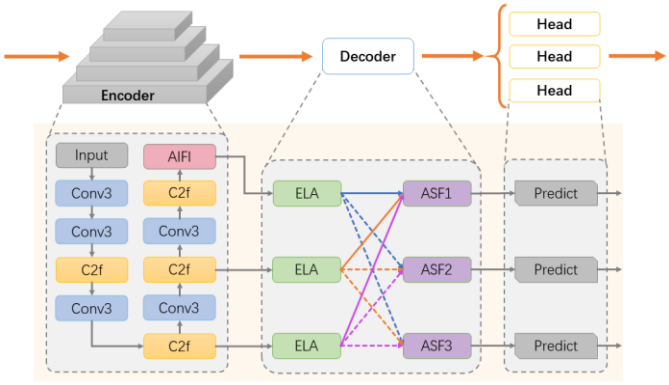

Fig. 2. The architecture of ELANet.

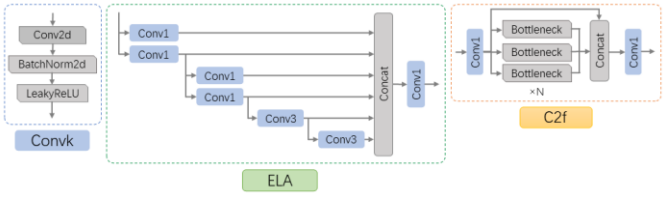

Fig. 3. Details of module design.

B. Encoder

The role of the encoder is to map the input RGB images into feature maps. Specifically, includes five downsampling operations performed by convolutional layers with a stride of 2 and a 3×3 kernel size. In the middle, a C2f module [23] is inserted for feature extraction, generating feature maps at different stages. The sequence of operations can be described as follows: Conv3-Conv3-C2f-Conv3-C2f-Conv3-C2f-Conv3-C2f (In Convk, 'k' represents the size of the convolution kernel). These operations are connected sequentially, with the

output of the previous layer serving as the input for the next layer. The initial channel number is 3, corresponding to the RGB image channels, and it increases gradually. It's noteworthy that in the final stage of the Encoder, a Multi-Head Self-Attention module called Attention-based Intra-scale Feature Interaction (AIFI) [24] is employed to handle the highest-level features of the backbone network, as depicted in Fig. 3. The mathematical processes are defined by Eq. (1) and Eq. (2):

$$Q = K = V = Flatten(S_{Last}) \quad (1)$$

$$F_{Last} = Reshape(Attn(Q, K, V)) \quad (2)$$

where, S_{Last} represents the last layer feature map output by the Encoder. Initially, the two-dimensional feature S_{Last} is flattened into a vector, which is then processed by the AIFI module. Subsequently, the output is reshaped back into two dimensions, denoted as F_{Last}, facilitating its transmission to the Decoder for feature analysis.

C. Decoder

The role of the Decoder is to combine and utilize features from the Encoder and decode predictive information. In visual models for processing remote sensing images, the utilization of gradient information is crucial. This process is generally achieved through continuous upsampling to facilitate the fusion of semantic information. Despite being a common approach, there is inevitably a problem of feature information loss or degradation, impacting the fusion effectiveness across non-adjacent levels. To address this issue, this study introduces the ELA module, whose structure is depicted in Fig. 3. Specifically, the ELA module employs two consecutive convolutional operations to capture richer feature representations. Subsequently, by introducing additional gradient flow branches in parallel, incorporates a broader context to enhance abstraction capabilities for targets. The features from the first six layers are concatenated to simultaneously utilize multi-scale information, strengthening the detection capabilities for targets of different sizes. Finally, a 1×1 convolutional layer is applied to reduce the dimensionality of the matrix, thereby alleviating the computational load of the model. With these connection operations, the model gains richer gradient information, achieving higher accuracy and more reasonable latency. Notably, in the ELANet model, a further adaptation is made using Adaptive Spatial Fusion (ASF) proposed by Yang et al. [25], which supports direct interaction between non-adjacent levels for adaptive spatial feature fusion. Through two consecutive convolutional operations, the ELA module adapts well to objects of different scales. The first convolutional operation captures features at a smaller scale, while the second convolutional operation integrates these features within a larger receptive field, making the model more flexible and accurate in detecting objects of different sizes. The design of the Decoder section integrates features from three different levels of the Encoder. ASF allocates different spatial weights to enhance the importance of key levels and reduce the impact of conflicting information from different levels.

Representation learning in convolutional neural networks faces the challenge of strong correlations between adjacent pixels, implying the potential provision of redundant information. To address this issue, there are some carefully designed strategies within each output feature of ASF, utilizing regularization techniques to alleviate domain variance. Specifically, a dropout strategy with a probability of 0.1 and Shuffle Attention (SA) attention strategy [26] during the upsampling and downsampling processes are employed. Since the zeroing of elements after dropout is random, connecting an SA attention strategy for responsive feature weight adjustment is deemed necessary. This necessity will be validated and analyzed in the experiments below. Finally, ELANet merges the decoded feature maps from different stages into the original feature image. Through pixel-level prediction, the regression branch predicts the distance from each anchor point to the four edges of the target bounding box, determining the target's position.

D. Loss Function

In the implementation of the ELANet model, use three loss functions to guide the regression of bounding boxes. The Classification Loss is used to measure the difference between the predicted class and the true class. This process is guided by the cross-entropy loss function, a common binary classification loss function, defined as follows:

$$L_c = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log p_i + (1 - y_i)\log(1 - p_i)] \quad (3)$$

In Eq. (3), let the ground truth be denoted as y, the predicted result as y, and n represents the batch size.

The Regression Loss is composed of Complete Intersection over Union (CIoU) and Distribution Focal Loss (DFL). CIoU is used to guide the model in learning the matching degree of bounding boxes. Specifically, assuming the predicted box and the ground truth box are denoted as b_p and b_g respectively, it is described as follows:

$$L_{CIoU} = IoU - \frac{\rho^2(b_p, b_g)}{c^2} - \alpha v \quad (4)$$

In Eq. (4), IoU represents the Intersection over Union, and $\rho^2(b_p, b_g)$ calculates the Euclidean distance between the center points of the two rectangular boxes. Here, c is the length of the diagonal of the minimum bounding rectangle of the predicted and ground truth boxes, v is used to measure the similarity of aspect ratios, and α is the impact factor of v. Next, dfl optimizes the position of the bounding boxes through smooth L1 loss. For each positive sample i, it is defined as:

$$L_{dfl} = \frac{1}{N_{pos}}\sum_{i=1}^{N_{pos}} Sml(pd_i, gt_i) \times w_i \quad (5)$$

In Eq. (5), N_{pos} represents the number of positive samples, $Sml(pd_i,\ gt_i)$ represents the smooth L1 loss for the i^{th} positive sample, and w_i is the weight associated with the i^{th} sample. Combining Eq. (4) and Eq. (5), the regression loss Lr can be obtained as $L_r = L_{CIoU} + L_{dfl}$.

The final loss for ELANet is defined as the weighted sum of the classification loss and regression loss, i.e., $L_{ELANet} = \alpha L_c + \beta L_r$.

III. EXPERIMENT

A. Experimental Details

1) Data preprocessing: The RSOD and UCAS-AOD datasets consist of 446 and 1000 airplane images, respectively, designated as the source domain and target domain. For convenience, this configuration is named RSOD. Beyond that, the GWHD dataset was intentionally set up for cross-domain settings, with 165 wheat spike images in the source domain and 71 in the target domain. In Fig. 4, present the size information for each instance in these two datasets. The RSOD dataset exhibits significant size differences between source and target domain data, while the primary difference in the GWHD dataset originates from variations in the external and internal environments.

2) Training details: The experiments are implemented in PyTorch [27] and accelerated using an NVIDIA RTX 3090 GPU. To improve computational efficiency, the longest side of input images is scaled to 608 pixels, and the other side is scaled proportionally, which also suits the resolution requirements when deploying on low-end edge devices. The Adaptive Moment Estimation (Adam) algorithm [28] was employed as the optimizer with a momentum factor set to 0.937, and the initial learning rate was set to 0.01. Considering convergence speed, perform 150 epochs of optimization on the RSOD dataset and 300 epochs on the GWHD dataset. To ensure the robustness of model training, strategies such as color distortion, random scale transformations, and mosaic data augmentation are employed.

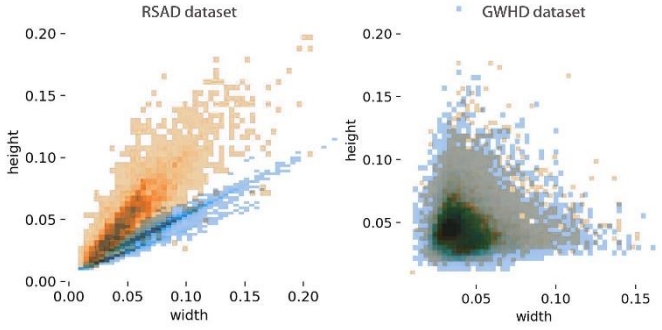

Fig. 4. Instance size information for each dataset.

B. Evaluation Indicators

In the process of establishing a detection model, it is essential to consider both precision and recall. Therefore, this study adopts metrics such as Precision, Recall, mAP@0.5, and mAP@0.5-0.95 to evaluate the model's performance and assess the detection results. The calculation methods for Precision and Recall are as Eq. (6) and Eq. (7):

$$P = \frac{TP}{TP+FP} \tag{6}$$

$$R = \frac{TP}{TP+FN} \tag{7}$$

where, P represents precision, and R represents recall. True Positive (TP) is the number of positive samples correctly classified, True Negative (TN) is the number of negative samples correctly classified, False Positive (FP) is the number of negative samples incorrectly classified as positive, and False Negative (FN) is the number of positive samples incorrectly classified as negative.

mAP represents the comprehensive performance at different Intersection over Union (IoU) thresholds, including mAP@0.5 and mAP@0.5-0.95. Here, mAP@0.5 denotes the average mAP when the IoU threshold is 0.5, with a higher value indicating higher detection precision for that category. mAP@0.5-0.95 represents the average mAP at different IoU thresholds (ranging from 0.5 to 0.95 with a step size of 0.05), placing stricter demands on the model's performance. The calculation of mAP is given by Eq. (8):

$$mAP = \frac{1}{n}\sum_{1}^{n} P(R)d(R) \tag{8}$$

where, n is the number of categories. In this experiment, each dataset has only one label, so n=1.

Furthermore, three metrics will be employed in the counting evaluation to assess the consistency between predicted and ground truth values, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2). Specifically, they are defined by Eq. (9) to Eq. (11):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{9}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{10}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\bar{y}_i - y_i)^2} \tag{11}$$

In these formulas, the numerator represents the sum of squared differences between the actual values and predicted values, and the denominator represents the sum of squared differences between the actual values and the mean. The result of R^2 falls within the range of [0, 1], indicating the proportion of the squared differences of predicted values to the squared differences of actual values near the mean. This metric can be understood as a measure of how well the model's predictions fit the actual values, with 1 indicating a perfect fit and 0 indicating no linear relationship between actual counts and predicted values.

C. Comparision with other Methods

To validate the effectiveness of ELANet, comparisons with two state-of-the-art methods: the two-stage model Faster R-CNN [29] and the one-stage model YOLOv7-tiny [30], both of which are widely used for visual tasks in images. Table I and Table II present the quantitative results on the two datasets, while Fig. 5 showcases some prediction examples from ELANet.

One notable observation is that, even though both datasets explicitly implement cross-domain settings, the performance on RSOD significantly surpasses that on GWHD. In reality, domain differences in the agricultural domain are extremely complex. The chaotic background makes the visual patterns of plants diverse and misleading, and the changes in plants themselves are also very pronounced. As shown in Fig. 1(a),

despite the small distribution difference between the source and target domains, the recognition performance is not ideal due to the confusion of different category samples. As demonstrated in Fig. 5(a), (b), the morphological differences in wheat spikes from different regions and varieties are substantial.

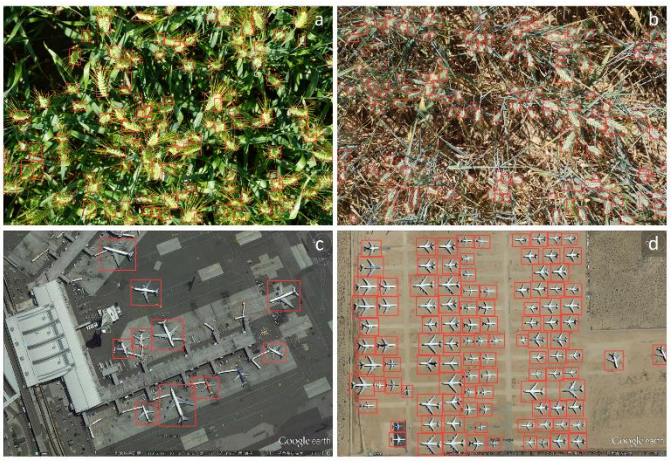

Fig. 5. Part of ELANet's prediction results. (a, b) are from the GWHD dataset, and (c, d) are from the RSOD dataset.

The comparison of object detection methods on the GWHD and RSOD datasets reveals distinct performance characteristics. According to the quantitative results shown in Table I for models on the GWHD dataset, ELANet achieves a precision of 88.7% without the need for adaptation to distribution differences on low-resolution input images. YOLOv7-tiny, while effective in terms of parameters, exhibits moderate precision and recall, with a lower mAP@0.5-0.95 score. Faster R-CNN, with a higher parameter count, shows performance comparable to YOLOv7-tiny but lacks the precision and recall achieved by ELANet, possibly due to information loss resulting from the simplification of their network structures. ELANet achieves optimal performance at the same input size of 608×608, substantially improving precision and average precision.

TABLE I. QUANTITATIVE RESULTS OF GWHD DATASET

Method	P	R	mAP@0.5	mAP@0.5-0.95	Params
YOLOv7-tiny	0.558	0.379	0.387	0.141	11.55M
Faster R-CNN	0.446	0.391	0.343	0.127	42.20M
ELANet	0.882	0.816	0.887	0.501	3.59M

Furthermore, as shown in Table II, the experimental results on the RSOD dataset also demonstrate a similar trend. However, the evaluation gap between the models is smaller in the RSOD dataset experiments. Additionally, it is worth noting that ELANet exhibits a smaller model parameter size, only 3.59M, compared to YOLOv7-tiny and Faster R-CNN with 11.55M and 42.2M, respectively. ELANet achieves significant improvements in both performance and parameter efficiency, indicating that it maintains high performance while being more parameter-efficient, making it well-suited for lightweight tasks without sacrificing efficiency.

TABLE II. QUANTITATIVE RESULTS OF RSOD DATASET

Method	P	R	mAP@0.5	mAP@0.5-0.95	Params
YOLOv7-tiny	0.555	0.821	0.802	0.310	11.55M
Faster R-CNN	0.832	0.751	0.753	0.278	42.20M
ELANet	0.894	0.845	0.861	0.337	3.59M

D. Analysis of Counting Influencing Factors

Due to the dense distribution of targets in both the GWHD and RSOD datasets, evaluating counting performance in this context is meaningful. Particularly in the case of the GWHD wheat head dataset, the model is prone to various natural factors during target detection, such as the influence of lighting, rainy weather, thick fog, etc. What's more, errors may arise from the varying shapes, sizes, and densities of different wheat head varieties. In this experiment, a linear regression plot was employed, along with counting metrics introduced in Section III (B), including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the Coefficient of Determination (R^2), to assess counting performance. The diagonal line on the coordinate system represents the ideal state where the model inference results perfectly match the manually counted ground truth. The linear regression plot serves as an effective tool to visually analyze the relationship between model predictions and actual counts, allowing researchers to gain a deeper understanding of the factors influencing counting performance. Additionally, it highlights images with the highest errors, as shown in Fig. 6.

It is evident that the maximum errors are primarily concentrated under varying lighting conditions, where inconsistencies in illumination affect wheat heads in bright and shadowed areas differently. The presence of cluttered foliage further challenges the target detection model. Through experimentation, it was discovered that even human experts find it challenging to discern in such conditions. Despite this, the ELANet model demonstrates good robustness, indicating that the optimized strategies of data augmentation play a positive role in enhancing the adaptability of the model for target detection performance. This also points towards future research directions, specifically addressing how to optimize models for strong interference environments to achieve stability, reliability, and adaptability in complex conditions.

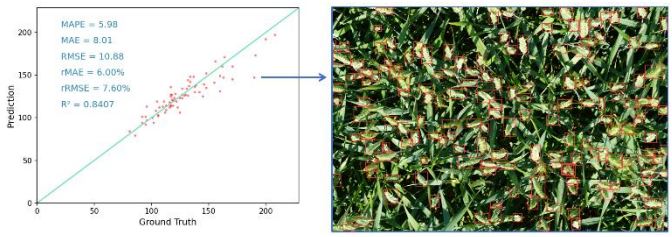

Fig. 6. Linear regression plot and maximum error plot of GWHD dataset count results.

E. Ablation Study

In the RSOD dataset, ELANet's performance is satisfactory. Building upon this, the focus shifted to conducting ablation experiments using Dropout strategy and SA strategy on the GWHD dataset. The experimental results are presented

in Table III and compared with two domain adaptation methods. In the table, "Dp" represents the use of the dropout strategy, and "At" represents the application of the SA attention strategy. It can be observed that without employing either strategy, ELANet achieves an accuracy of 0.87. However, using either the dropout strategy or attention strategy alone leads to a performance decrease in the detection task, with reductions of 1.72% and 0.23%, respectively. On the other hand, employing the SA attention strategy after using the dropout strategy proves to be highly effective. The random zeroing of elements after dropout reduces the risk of overfitting, preventing excessive co-adaptation of neurons and maintaining feature diversity. This contributes to improving the model's generalization performance and robustness. The SA attention strategy can re-weight features in response to enhance attention to important regions, adaptively fusing contextual information of different scales.

Another interesting finding is that when the training set and the validation set share the same settings, the model's performance is improved, as shown in Table IV, method A involves separate training and validation sets, while method B uses the same set for training and validation. From a subjective perspective, the model in this situation should not be robust. In reality, data transformation techniques alleviate the impact of this situation, and the performance improvement is attributed, to some extent, to the slightly increased training data.

TABLE III. ELANET USES DROPOUT STRATEGIES FOR ABLATION STUDIES

Dp	At	P	R	mAP@0.5	mAP@0.5-0.95
--	--	0.855	0.810	0.870	0.463
√	--	0.848	0.790	0.855	0.457
--	√	0.850	0.803	0.868	0.472
√	√	0.882	0.816	0.887	0.501

TABLE IV. COMPARISON OF ELANET PERFORMANCE WITH DIFFERENT DATASET SETTINGS

Method	P	R	mAP@0.5	mAP@0.5-0.95
A	0.872	0.779	0.864	0.468
B	0.882	0.816	0.887	0.501

IV. DISCUSSION AND FUTURE WORK

In this study, the challenges of cross-domain object detection were explored, focusing on addressing distribution differences between different datasets. The proposed ELANet model, based on feature extraction and fusion, was introduced. Experimental results indicate that ELANet performs exceptionally well on the RSOD dataset, while its performance on the GWHD dataset is relatively lower. This difference may be attributed to the clearer features in the airplane images of the RSOD dataset, making it easier for the model to learn effective feature representations. In contrast, the complex background and the similarity in color between target objects and the background in the GWHD dataset increase the difficulty of model recognition.

Furthermore, this study compared ELANet's performance on the RSOD and GWHD datasets with other methods, revealing superior performance on both datasets. This suggests that ELANet can better handle cross-domain challenges and improve the accuracy of object detection. Nevertheless, there are still some issues to be addressed. For example, ELANet's performance is influenced by data preprocessing and training details. To further enhance the model's performance, deeper research into data preprocessing and training techniques is needed. Distribution differences have been a long-standing issue in machine learning [31]-[35], and it is hoped that this work will further stimulate researchers' interest in addressing this problem.

In future research, potential improvement avenues can be explored in the following directions:

Adversarial Robustness Learning: Conduct in-depth research to assess ELANet's robustness against adversarial attacks. Strengthening the model's security is crucial for real-world deployment and addressing potential security challenges.

Testing in Complex Environments: Test ELANet's robustness in more complex environmental conditions, especially in scenarios with natural factors such as varying lighting and rainy weather. Consider employing more powerful data augmentation and processing techniques to enhance the model's adaptability to these challenges.

Driving Agricultural Technology Innovation: Expand ELANet's application to more agricultural domains, such as agricultural robots, precision agriculture, and plant disease diagnosis. Through widespread application in agricultural technology, ELANet aims to provide more intelligent and efficient solutions for agricultural production.

V. CONCLUSION

In this study, the focus was on exploring a visual model for domain adaptation without the need for explicit adaptation, particularly addressing the challenge of domain adaptation in supervised graph learning tasks. In this work, the ELANet model was proposed, innovatively introducing the ELA module and integrating it into the feature decoder, successfully achieving high-quality domain adaptation in the remote sensing image domain. The model demonstrated outstanding performance on real datasets. The research indicates that regularization techniques are crucial for mitigating the domain variance between training and testing data, especially when incorporating a module that reweights features in response. Importantly, the use of the ELANet model not only improved accuracy but also enhanced efficiency. In experiments validating the generality and domain adaptation of ELANet, cross-domain settings were employed, demonstrating the model's generality and domain adaptability. Overall, the introduction of ELANet is expected to advance research in domain adaptation, providing an innovative approach to addressing domain adaptation challenges in supervised graph learning tasks.

REFERENCES

[1] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 159, January 2020, pp. 296-307.

[2] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345-1359, Oct. 2010.

[3] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian Detection: An Evaluation of the State of the Art," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 4, pp. 743-761, April 2012.

[4] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 2012, pp. 2066-2073.

[5] Y. Ganin, E. Ustinova, H. Ajakan, et al., "Domain-adversarial training of neural networks," Journal of Machine Learning Research, vol. 17, no. 59, 2016, pp. 1-35.

[6] C. Doersch and A. Zisserman, "Multi-task Self-Supervised Visual Learning," in 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2070-2079.

[7] J. Casebeer, N. J. Bryan, and P. Smaragdis, "Meta-AF: Meta-Learning for Adaptive Filters," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, 2023, pp. 355–370.

[8] W. Wang, J. Dai, Z. Chen, et al., "Internimage: Exploring large-scale vision foundation models with deformable convolutions," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 14408-14419.

[9] P. Goel and A. Ganatra, "Unsupervised Domain Adaptation for Image Classification and Object Detection Using Guided Transfer Learning Approach and JS Divergence," Sensors, vol. 23, no. 9, 2023, pp. 4436.

[10] A. Elshamli, G. W. Taylor, A. Berg, et al., "Domain adaptation using representation learning for the classification of remote sensing images," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 10, no. 9, 2017, pp. 4198-4209.

[11] J. Zhang, S. Xu, J. Sun, et al., "Unsupervised Adversarial Domain Adaptation for Agricultural Land Extraction of Remote Sensing Images," Remote Sensing, vol. 14, no. 24, 2022, pp. 6298.

[12] M. Valerio Giuffrida, A. Dobrescu, P. Doerner, et al., "Leaf counting without annotations using adversarial unsupervised domain adaptation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 0-0.

[13] S. Marino, P. Beauseroy, A. Smolarz, "Unsupervised adversarial deep domain adaptation method for potato defects classification," Computers and Electronics in Agriculture, vol. 174, 2020, 105501.

[14] A. Fuentes, S. Yoon, T. Kim, et al., "Open set self and across domain adaptation for tomato disease recognition with deep learning techniques," Frontiers in Plant Science, vol. 12, 2021, 758027.

[15] X. Wu, X. Fan, P. Luo, et al., "From Laboratory to Field: Unsupervised Domain Adaptation for Plant Disease Recognition in the Wild," Plant Phenomics, vol. 5, 2023, 0038.

[16] F. Magistri, J. Weyler, D. Gogoll, et al., "From one field to another—Unsupervised domain adaptation for semantic segmentation in agricultural robotics," Computers and Electronics in Agriculture, vol. 212, 2023, 108114.

[17] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition," in 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 2014, pp. 512-519.

[18] H. Lu, C. Shen, Z. Cao, Y. Xiao, and A. van den Hengel, "An Embarrassingly Simple Approach to Visual Domain Adaptation," IEEE Transactions on Image Processing, vol. 27, no. 7, 2018, pp. 3403–3417.

[19] E. David, M. Serouart, D. Smith, et al., "Global Wheat Head Detection 2021: An Improved Dataset for Benchmarking Wheat Head Detection Methods," Plant Phenomics, Sep 22, 2021;2021:9846158.

[20] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate Object Localization in Remote Sensing Images Based on Convolutional Neural Networks," in IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 5, May 2017, pp. 2486-2498.

[21] H. Zhu, X. Chen, W. Dai, K. Fu, Q. Ye, and J. Jiao, "Orientation Robust Object Detection in Aerial Images Using Deep Convolutional Neural Network," in 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 2015, pp. 3735-3739.

[22] C. Wang, H. Liao, and I. Yeh, "Designing Network Design Strategies Through Gradient Path Analysis," arXiv preprint arXiv:2211.04800, 2022.

[23] G. Jocher, A. Stoken, A. Chaurasia, J. Borovec, Y. Kwon, K. Michael, et al., "Yolov5 Repository," Available at https://github.com/ultralytics/yolov5, Accessed on 10/28/2023.

[24] W. Lv, S. Xu, Y. Zhao, et al., "Detrs beat yolos on real-time object detection," arXiv preprint arXiv:2304.08069, 2023.

[25] G. Yang, J. Lei, Z. Zhu, et al., "AFPN: Asymptotic Feature Pyramid Network for Object Detection," arXiv preprint arXiv:2306.15988, 2023.

[26] Q. L. Zhang and Y. B. Yang, "SA-Net: Shuffle Attention for Deep Convolutional Neural Networks," in ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 2235-2239.

[27] A. Paszke, S. Gross, F. Massa, et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," Advances in Neural Information Processing Systems, vol. 32, 2019.

[28] D.P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," arXiv preprint arXiv:1412.6980, 2014.

[29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, June 1, 2017.

[30] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors," in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 7464-7475.

[31] Q. Ming, L. Miao, Z. Zhou, and Y. Dong, "CFC-Net: A Critical Feature Capturing Network for Arbitrary-Oriented Object Detection in Remote-Sensing Images," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-14, 2022, Art no. 5605814.

[32] X. Hu and C. Zhu, "Shared-Weight-Based Multi-Dimensional Feature Alignment Network for Oriented Object Detection in Remote Sensing Imagery," Sensors, vol. 23, no. 1, 2022, Article 207.

[33] Q. Ming, L. Miao, Z. Zhou, J. Song, and X. Yang, "Sparse Label Assignment for Oriented Object Detection in Aerial Images," Remote Sensing, vol. 13, no. 14, 2021, Article 2664.

[34] S. Falahat and A. Karami, "Maize Tassel Detection and Counting Using a YOLOv5-Based Model," Multimedia Tools and Applications, vol. 82, pp. 19521–19538, 2023.

[35] Y. Yang and S. Soatto, "FDA: Fourier Domain Adaptation for Semantic Segmentation," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 4084-4094.