Towards Accurate Billboard Detection: An Ablation and Benchmarking Study of Deep Learning Models

Sukriti Dhang[1,3], Atseosi Idogho[1], Mimi Zhang[2], Soumyabrata Dev[1,3]

[1]*School of Computer Science, University College Dublin, Ireland*
[2]*School of Computer Science and Statistics, Trinity College Dublin, Ireland*
[3] *The ADAPT SFI Research Centre, Dublin, Ireland*

Abstract

With the increase in multimedia videos, the applications of videos and images to leverage targeted advertisements are becoming significantly diverse in reaching specific audiences. Targeted advertisements are achieved by detecting and replacing an existing advertisement within an image frame with a new advert. In this work, we classify image frames as advert or no-advert. We conducted a benchmarking study using two datasets: the AdShop dataset and the ALOS dataset. Our evaluation indicates that the AdShop dataset is suboptimal for billboard detection. Conversely, the ALOS dataset, specifically curated for billboard identification, demonstrates superior accuracy and reliability. We also performed an ablation study by reducing the filter size at different layers to estimate the classification performance. The results suggest that the proposed ModifiedVGGNet achieves the highest performance with a training accuracy of 97.2% and a test accuracy of 97.0%. This was achieved by reducing 10% filter in conv1 layer of block 1.

Keywords: Advert classification, deep learning, benchmark study, ablation study, k-fold cross validation

1 Introduction

Advertisements in multimedia videos play a crucial role in capturing the attention of a broad audience. Recent interest focuses on integrating advertising boards with modern techniques [Krishna and Aizawa, 2018], enhancing effectiveness in the evolving digital advertising landscape. The traditional approaches [Nautiyal et al., 2019] not only benefit the effectiveness of advertising, but also contribute to efficient urban planning, optimizing space, and reducing visual clutter. In the present years, there has been significant progress in the field of object detection and semantic segmentation, with methodologies to classify and localize objects in visual scenes [St-Charles et al., 2014]. However, the work involving the general task of classifying billboards/advertisement boards in outdoor scenes is limited. Detecting billboards in a video sequence involves identifying them within each image frame, which is crucial for digital advertising [Yu et al., 2023]. In the traditional approach, video editors manually inspect the video frames to incorporate new advertisements which requires a significant amount of time and labor. The efficient and accurate detection of billboards is essential for seamless integration of new advertisements into existing frames. Our approach employs state-of-the-art methodologies from deep learning to effectively detect advertisements in the image frames as either advert (class ‘0’) or no-advert (class ‘1’).

In this work, we focus on evaluating the accuracy of advert classification within image frames, by proposing an automated deep learning model based on an ablation study. To develop the model, we fine-tuned the performance of existing state-of-the-art models by freezing the weights of the initial layers. These models were then compared to ascertain the best one based on accuracy. An ablation study is conducted to the best model (VGG19) which achieves the highest accuracy by reducing the filter size to increase the model’s robustness. A k-fold cross-validation [Wong and Yeh, 2020] is performed to ensure that overfitting does not occur. By partitioning the dataset into k-folds, with each fold can be used for both training and validation, iteratively through

https://doi.org/10.1049/icp.2024.3303
© This is an open access article published by the IET under the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/)

different combinations until each subset has been utilized for validation at least once. This process allows for a more robust assessment of model performance.

The main objective of this work is to construct an automated deep learning model to assist video editors, including movie making, advertising agencies, and sports broadcasting by classifying image frames into advert and no-advert with accurate performance. The main contributions to this work [1], are as follows: *i*) Pre-trained networks such as VGGNet19 and ResNet are adapted by fine-tuning the final layers of the model to identify the most effective classifier for the dataset. *ii*) An ablation study is conducted on the proposed ModifiedVGGNet model to enhance its classification performance. *iii*) Proposed ModifiedVGGNet leveraged a fine-tuned VGG19 architecture as it shows improved accuracy compared to other state-of-the-art models. *iv*) To mitigate the risk of overfitting, the model is evaluated using the k-fold cross-validation technique.

The rest of the paper is organized as follows: Section 2 provides the background and related work. In Section 3, we discuss the dataset utilized in this study and outline the proposed methodology. Section 4 presents a comparative analysis using the benchmarking dataset. Finally, Section 5 concludes the paper by summarizing the results.

2 State of the Art

Over the past few years, deep convolutional neural networks have proven significant effectiveness in image and video processing domains. Particularly, they have found an extensive application in addressing challenges related to image classification and object detection [Girshick et al., 2014]. The existing work in advert detection mainly focuses on identifying advertisement clips within a video sequence.

Recent advancements in computer vision and machine learning fields, billboard/advertisement classification has been a subject of research, and various methods have been employed to tackle this problem. Chavan *et.al* [Chavan et al., 2021] proposed a method for billboard location by manually annotating a bounding box and training a convolutional neural network to automatically recognize the billboard. The proposed model predicts bounding boxes and the classes of advertisement billboards with 60 % accuracy. Rahmat *et.al.* [Rahmat et al., 2019] retrained a pre-trained Inceptionv3 model with an additional 384 billboard images. However, the model performance was poor in detecting billboards under low light conditions.

Several studies suggested to perform k-fold cross-validation repeatedly to obtain reliable accuracy estimates for statistical comparison. Kohavi [Kohavi, 1995] recommended that the number of folds used for variance estimation should not be less than ten. Vanwinckelen and Blockeel [Vanwinckelen and Blockeel, 2012] and Kim [Kim, 2009] observed that the repetition of k-fold cross-validation helps in stabilizing the variability of accuracy estimates. Demšar [Demšar, 2006] introduced nonparametric methods for evaluating the performance of classification algorithms, which depend on dependable estimates acquired through repeated stratified random splits. Therefore, it is interesting for us to explore k-fold cross validation techniques to make accurate classification and perform ablation study to increase model's robustness.

3 Methodology

We experimented with VGG19Net and Res50Net models to determine the optimal network in terms of accuracy. We fine-tuned the model by adding one flatten layer and two dense layers to the base of the pre-trained models, and modified the output layer to have two neurons.

3.1 Dataset

AdShop dataset This Adshop dataset [YusukeKumakoshi, 2021] consists of sign boards and the shop names along with specific annotations or categories related to billboards/advertisements labeled as billboard in Japanese

[1] To facilitate the reproducibility of this research, the code of this paper is made available: `https://github.com/sukritidhang/billboard-detect-ablationstudy`

Streetscapes images along with corresponding masks. This corresponds to 1750 images as the positive class (images containing an advertisement) with resolution of 480 × 360. The billboards in the images included were of a variety of heights, orientations and shapes. The dataset's size remains limited, which presents challenges for comprehensive analysis and precise predictions in tasks like classification.

ALOS dataset For benchmarking purposes, we utilized the ALOS dataset [Dev et al., 2019], which comprises 9,315 images with a resolution of 512 × 512. Prior to this, there exists no publicly available dataset with manually annotated billboard maps. To train the models, we chose 1,750 images as the positive class (images containing an advertisement). This ALOS dataset ensures the inclusion of diverse billboard characteristics, with billboards occupying varying proportions of the total image area. We selected 1750 images as negative class (images without any advertisement) [Heonho, 2022][Bansal, 2019]. Images in both classes cover a wide variety of scenarios and landscapes in order to train the model, such as urban and rural areas, day- and night-time scenes. The description of the dataset is summarized in Table 1.

Table 1: Dataset description

Name	**Description**	
	AdShop Dataset	**ALOS Dataset**
Total number of images	3500	9315
Dimension	480 x 360	512 x 512
Color grading	RGB	RGB
Advert images	1750	1750
No-advert images	1750	1750

3.2 Model Framework

In traditional machine learning approaches, Support Vector Machines (SVMs) [Cortes and Vapnik, 1995] and Random Forest [Ho, 1995] can be used for image classification by representing images as feature vectors. In this work, we used RBF kernel, a C value of 0.4233 for regularization, and a gamma value of 0.01. However, training the model for 100 iterations was insufficient for convergence, highlighting the computational expense of SVMs. To address these limitations and improve performance, a deep learning based simple model has been created with three convolutional layer and max pooling layer pairs, two of which have their filter size, kernel size, and strides length chosen using the hyper-parameter tuning function. Followed by a dense layer with a filter size of 1024, a dropout layer with a rate of 0.5, and a final output dense layer has 2 neurons with a soft-max activation function. The sparse categorical cross-entropy loss function is used to calculate the loss, and Adam optimiser with a hyper-parameter tuned learning rate of 0.00225. The Simple model has been trained from scratch using the datasets. As our dataset is limited to improve accuracy, we incorporated AdNet [Hossari et al., 2018] which uses pre-trained weights of VGG19 a convolutional neural network (CNN) architecture that consists of 19 layers, including 16 convolutional layers, three fully connected layers, and five max pooling layers. To prevent overfitting in ADNet, a dropout layer with a rate of 0.5 was introduced. Additionally, the second fully connected (FC) layer consists of 1024 channels with a ReLU activation function. The final FC layer comprises 2 neurons with a softmax activation function. These two output values from ADNet represent the probabilities of belonging to the advert or no-advert classes. For benchmarking purpose, we utilize ResNet [He et al., 2016] a CNN architecture that uses residual blocks as the pre-trained model with the 50-layers. The head follows the structure of the ADNet model, including a dropout layer with a rate of 0.5. The second fully connected (FC) layer also consists of 1024 channels, and the final FC layer comprises 2 neurons with a ReLU activation function.

In our work, we selected the best model based on accuracy. Then, we employ k-fold cross-validation techniques to improve the accuracy of our model. We have partitioned the dataset into 10 equal folds. Subsequently, the model is trained and validated ten times, each fold being designated as the validation set once, while the remaining folds are used for training. By rotating the validation set across all folds, we ensure comprehensive

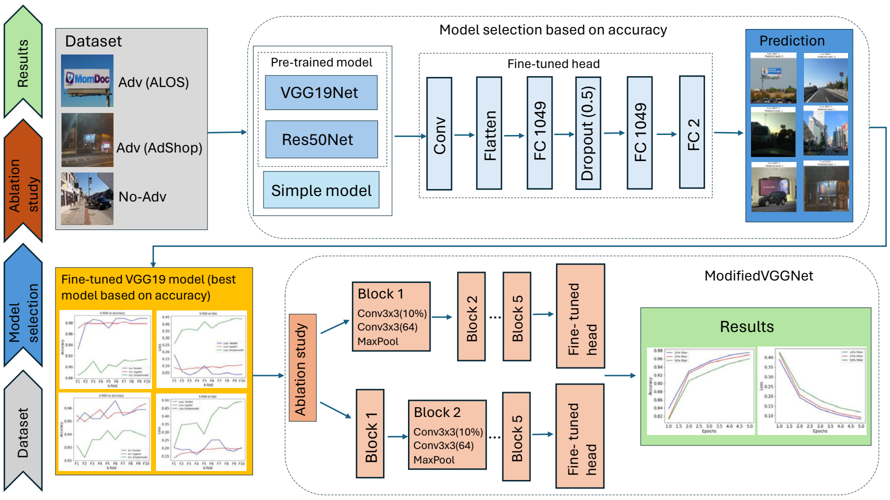


Figure 1: Workflow of the billboard detection.

model assessment. We analyze the prediction results from various k-fold cross-validations to show that the accuracy estimates obtained are dependent. We also conducted an ablation study on the best model (VGG19) by systematically reducing the filter size to 10%, 25%, and 50% to further increase the model's robustness. By systematically adjusting the filter sizes, we aimed to understand how this modification influenced the model's classification performance. The entire framework of advert classification proposed in this research is illustrated in Figure 1. Detailed explanation of the results of our ablation study on ModifiedVGGNet is given in the section 4.2.

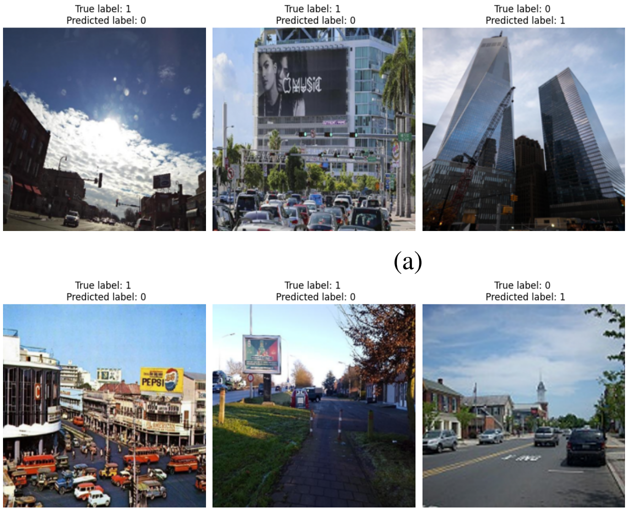
4 Experimental results

The classification of the image is predicted as either Label '1' or Label '0', where Label '1' resembles a positive class that belongs to the Advert class and Label '0' resembles negative class that belongs to the No-advert class. When utilizing the AdShop dataset for detection tasks, the performance is hindered due to inconsistent labeling of the data. A common issue is the misclassification of objects, such as shop nameplates being incorrectly labeled as adverts. This inconsistency confuses the detection algorithms, leading to lower accuracy and higher error rates in classifying the advertisements correctly. The Figure 2 and Figure 3 depicts the misclassification as incorrectly identified as advert and no-advert. On the other hand, the ALOS dataset is much more reliable in terms of labeling accuracy. The labels in the ALOS dataset are consistent and correctly assigned to the corresponding objects. This accurate labeling allows detection algorithms to perform better, as the training data is more representative of the actual objects they need to detect. Consequently, detection performance is significantly enhanced when the ALOS dataset compared to the AdShop dataset.

4.1 Performance Metrics

In this work, 10-fold cross-validation is employed to evaluate the performance of a classification model on a dataset that contains 1,750 instances in each of two classes Advert and No-advert classes. The experimental results are shown in the Table 2 represents the training results for 10 k-folds. The performance evaluation of our model employing K-fold cross-validation is illustrated in Figure 4. Within the K-fold cross-validation algorithm, one of the k folds is designated as the validation set V, while the remaining folds serve as the training set. The validation process assesses the accuracy as (Acc_k) using the equation (1).

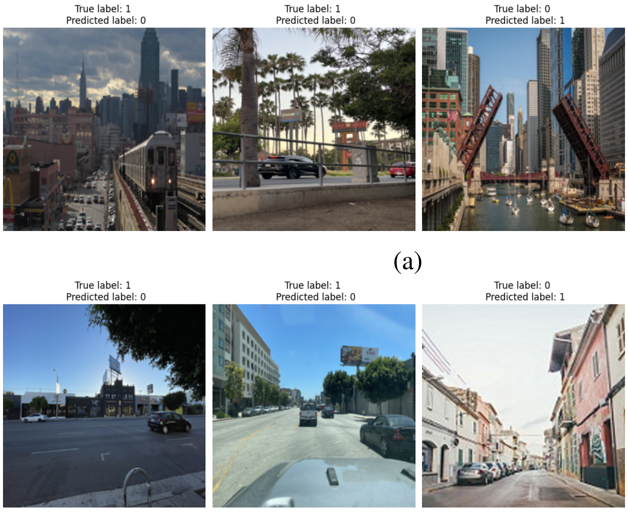
$$Acc_k = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$



(a)

(b)

Figure 2: We represent experimental result of misclassifiation as incorrectly identified as advert and no-advert (a) results of VggNet, (b) results of ResNet, for ALOS dataset.



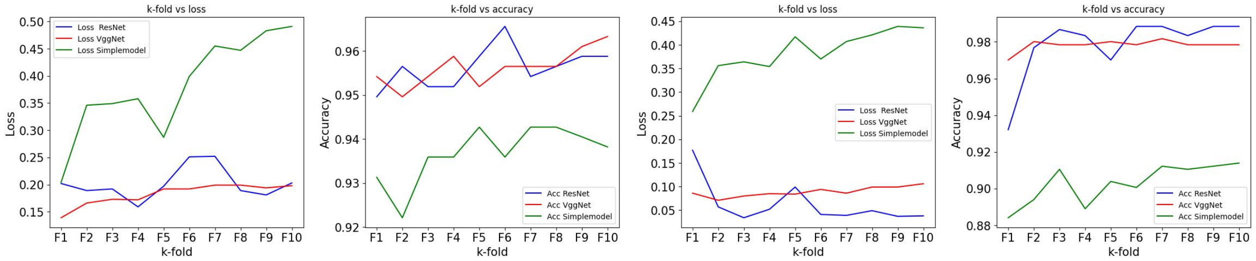
(a)

(b)

Figure 3: We represent experimental result of misclassifiation as incorrectly identified as advert and no-advert (a) results of VggNet, (b) results of ResNet for AdShop dataset.

Table 2: Training results of Vgg19Net, Res50Net, Simple model across 10 folds before performing abalation study.

Model	**AdShop dataset**		**ALOS dataset**	
	Avg. Accuracy	**Loss**	**Avg. Accuracy**	**Loss**
Vgg19Net	95.62 (± 0.38)	0.182	97.83 (± 0.29)	0.089
Res50Net	95.62 (± 0.43)	0.200	97.86 (± 1.65)	0.063
Simple model	93.68 (± 0.60)	0.382	90.31 (± 1.02)	0.382



(a)

(b)

(c)

(d)

Figure 4: Representation of performance measurement of the models using k-fold cross-validation before performing ablation study. We represent trend of 10-fold (a) loss curve and (b) accuracy curve for AdShop dataset, (c) loss curve and (d) accuracy curve for ALOS dataset.

Here, TP represents true positives, TN stands for true negatives, FP indicates false positives, and FN signifies false negatives. The number of Acc_k values will correspond to the value of k. The final step is to aggregate these values using equation (2), where Acc is aggregate accuracy.

$$Acc = \frac{1}{K}\sum_{k=1}^{K} Acc_k \tag{2}$$

The Table 3 represents the testing results of different models on two datasets: AdShop and ALOS. We calculated the metrics including accuracy, the number of misclassified instances, precision, recall, F1-score, and AUC (Area Under the ROC Curve). From the result, VggNet slightly outperforms ResNet with high accuracy and strong performance across all metrics on both datasets. Simple Model has a lower accuracy on the AdShop dataset compared to Vgg19Net and Res50Net, but performs well on the ALOS dataset, particularly in precision and recall. Random Forest has decent performance on the AdShop dataset but shows significantly lower accuracy and other metrics on the ALOS dataset. SVM underperforms significantly on both datasets with very low accuracy and high misclassification.

Table 3: Testing results of VggNet, ResNet, Simple model before performing ablation study. Total testing images in AdShop dataset are 437 images and 604 images in ALOS dataset.

Model	**AdShop dataset**						**ALOS dataset**					
	Accuracy	**Misclass**	**Precision**	**Recall**	**F1-score**	**AUC**	**Accuracy**	**Misclass**	**Precision**	**Recall**	**F1-score**	**AUC**
Vgg19Net	**0.96**	**15**	**0.969**	**0.972**	**0.971**	**0.964**	**0.98**	**8**	**0.984**	**0.990**	**0.987**	**0.986**
Res50Net	0.96	17	0.976	0.957	0.966	0.961	0.98	8	0.987	0.987	0.987	0.986
Simple Model	0.94	23	0.983	0.926	0.954	0.952	0.92	50	0.987	0.987	0.987	0.915
Random forest	0.91	31	0.940	0.911	0.925	0.913	0.79	113	0.757	0.900	0.821	0.791
SVM	0.80	85	0.810	0.876	0.842	0.789	0.56	267	0.547	0.987	0.700	0.532

4.2 Ablation study

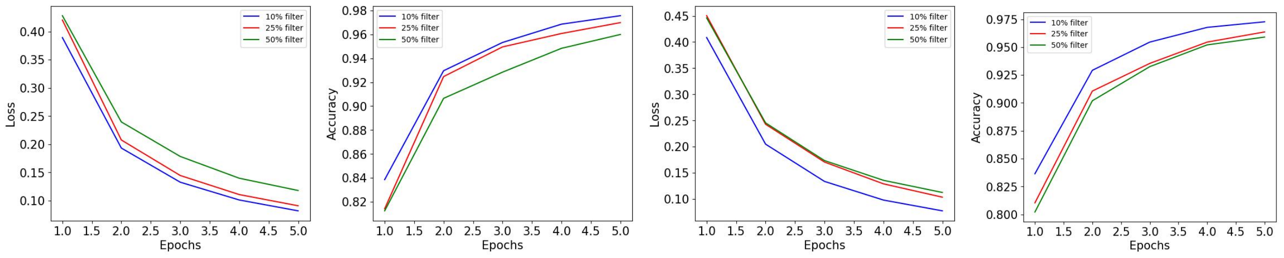
After fine-tuning and selecting the best model based on accuracy, we performed an ablation study as presented in Table 4 to further increase the robustness of the model. Figure 5 shows the accuracy for iteratively performed ablations of 10%, 25% and 50% of filters in conv1 of block 1 and block2 of VGG-19. The impact of filter reduction on the model's performance is nuanced. However, these models eventually achieve comparable or better validation accuracy, suggesting improved generalization. For Block 1 conv 1 (B_1C_1), the accuracy remains identical to the baseline, achieving high performance with slight overfitting. Block 2 conv 1 (B_2C_1), shows increased accuracy, demonstrating improved generalization and lower test loss, indicating the best performance among the filter sizes tested. At a 25% filter size, B_1C_1 maintains identical accuracy with consistent performance across different configurations, while B_2C_1 experiences a drop in accuracy, reflected by a slight decrease in test accuracy and higher test loss. However, with a 50% filter size, for B_1C_1 the model exhibits an increase in accuracy , whereas with 50% filter size in B_2C_1 shows a drop in accuracy. Overall, the 25% filter size provides stable performance but can lead to a decrease in accuracy, while the 50% filter size yields variable results, improving accuracy in some scenarios but declining in others. Therefore, from the result, 10% filter size, especially in B_2C_1, achieve optimal accuracy and generalization. For future work, it would be beneficial to investigate further reductions in other blocks to determine whether similar trends hold. Additionally, exploring different training setups, such as longer training duration's or using different datasets, could help validate these findings.

Table 4: Representation of ablation study on the proposed ModifiedVGGNet model by reducing filter size.

Case Study	**Filter size**	**Layer**	**Train loss**	**Train acc**	**Test loss**	**Test acc**	**Precision**	**Recall**	**F1-score**	**RMSE**	**AUC**	**Finding**
1	10%	B_1C_1	0.081	0.975	0.111	0.958	0.945	0.978	0.961	0.203	0.957	Identical accuracy
		B_2C_1	0.077	0.972	0.094	0.970	0.968	0.974	0.971	0.172	0.969	Accuracy increased
2	25%	B_1C_1	0.090	0.969	0.107	0.958	0.950	0.971	0.961	0.203	0.957	Identical accuracy
		B_2C_1	0.103	0.963	0.129	0.953	0.953	0.959	0.956	0.215	0.953	Accuracy dropped
3	50%	B_1C_1	0.118	0.959	0.118	0.963	0.943	0.990	0.966	0.190	0.961	Accuracy increased
		B_2C_1	0.112	0.959	0.132	0.951	0.934	0.978	0.955	0.219	0.950	Accuracy dropped

5 Conclusion

In this work, we performed an ablation and benchmark the model using two datasets: the AdShop dataset and the ALOS dataset. Our evaluation reveals that the AdShop dataset is suboptimal for billboard detection, as it



(a) (b) (c) (d)

Figure 5: Representation of ablation study results with 10%, 25% and 50% of filters reductions: (a) loss curves for Block 1 Conv1, (b) accuracy for Block 1 Conv1, (c) loss curves for Block 2 Conv1, and (d) accuracy for Block 2 Conv1 of the ModifiedVGGNet model.

contains misclassified labels, notably where shop names are incorrectly labeled as billboards. This inconsistency confuses the detection algorithms, resulting in lower accuracy and higher error rates. Conversely, the ALOS dataset, specifically curated for billboard identification, demonstrates superior accuracy and reliability. The results underscore the importance of the quality of the dataset in achieving accurate benchmarking outcomes in frame-level localization tasks. Therefore, selecting domain-specific high-quality datasets is crucial for reliable performance assessment in similar applications. An ablation study is performed to the best model (VGG19), which yields the highest accuracy by reducing the filter size to further increase the model's robustness. The results suggest that the proposed ModifiedVGGNet performs best with a training accuracy of 97.2%, and a test accuracy of 97.0% after reducing 10% filter in the conv1 layer of block 1.

Acknowledgments

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT SFI Research Centre at University College Dublin. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, is funded by Science Foundation Ireland through the SFI Research Centres Programme.

References

[Bansal, 2019] Bansal, P. (2019). Intel Image Classification. `https://www.kaggle.com/datasets/puneet6060/intel-image-classification`.

[Chavan et al., 2021] Chavan, S., Kerr, D., Coleman, S., and Khader, H. (2021). Billboard detection in the wild. In *Irish Machine Vision and Image Processing Conference 2021*, pages 57 – 64. Irish Pattern Recognition and Classification Society.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

[Demšar, 2006] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7:1–30.

[Dev et al., 2019] Dev, S., Hossari, M., Nicholson, M., McCabe, K., Conran, A. N. C., Tang, J., Xu, W., and Pitié, F. (2019). The ALOS dataset for advert localization in outdoor scenes. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE.

[Girshick et al., 2014] Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[Heonho, 2022] Heonho (2022). Unpaired Day and Night cityview images. `https://www.kaggle.com/datasets/heonh0/daynight-cityview`.

[Ho, 1995] Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE.

[Hossari et al., 2018] Hossari, M., Dev, S., Nicholson, M., McCabe, K., Nautiyal, A., Conran, C., Tang, J., Xu, W., and Pitié, F. (2018). ADNet: A deep network for detecting adverts. *arXiv preprint arXiv:1811.04115*.

[Kim, 2009] Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational statistics & data analysis*, 53(11):3735–3745.

[Kohavi, 1995] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada.

[Krishna and Aizawa, 2018] Krishna, O. and Aizawa, K. (2018). Billboard saliency detection in street videos for adults and elderly. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2326–2330.

[Nautiyal et al., 2019] Nautiyal, A., McCabe, K., Hossari, M., Dev, S., Nicholson, M., Conran, C., McKibben, D., Tang, J., Xu, W., and Pitié, F. (2019). An advert creation system for next-gen publicity. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part III 18*, pages 663–667. Springer.

[Rahmat et al., 2019] Rahmat, R. F., Dennis, D., Sitompul, O. S., Purnamawati, S., and R., B. (2019). Advertisement billboard detection and geotagging system with inductive transfer learning in deep convolutional neural network. *Telkomnika (Telecommunication Computing Electronics and Control)*, 17(5):2659 – 2666.

[St-Charles et al., 2014] St-Charles, P.-L., Bilodeau, G.-A., and Bergevin, R. (2014). Subsense: A universal change detection method with local adaptive sensitivity. *IEEE Transactions on Image Processing*, 24(1):359–373.

[Vanwinckelen and Blockeel, 2012] Vanwinckelen, G. and Blockeel, H. (2012). On estimating model accuracy with repeated cross-validation. In *BeneLearn 2012: Proceedings of the 21st Belgian-Dutch conference on machine learning*, pages 39–44.

[Wong and Yeh, 2020] Wong, T.-T. and Yeh, P.-Y. (2020). Reliable accuracy estimates from k-fold cross validation. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1586–1594.

[Yu et al., 2023] Yu, L., Li, G., Yuan, L., and Zhang, L. (2023). Time-bounded targeted influence spread in online social networks. *Multimedia Tools and Applications*, 82(6):9065–9081.

[YusukeKumakoshi, 2021] YusukeKumakoshi (2021). Billboard in japanese streetscapes. `https://www.kaggle.com/datasets/yusukekumakoshi/billboard-in-japanese-streetscapes`.