Machine Learning with Applications 21 (2025) 100686

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

AdVision: An efficient and effective deep learning based advertisement detector for printed media

Faeze Zakaryapour Sayyad [a,b,*], Irida Shallari [b], Seyed Jalaleddin Mousavirad [b], Mattias O'Nils [b], Faisal Z. Qureshi [c]

[a] Media Research AB, Ringvägen 31, Östersund, 831 37, Sweden
[b] Mid Sweden University, Holmgatan, Sundsvall, 852 30, Sweden
[c] Faculty of Science, University of Ontario Institute of Technology, 2000 Simcoe Street North, Oshawa, Ontario L1G 0C5, Canada

ARTICLE INFO

Keywords:
Cross-linguistic advertisement detection
Deep learning
Newspaper image analysis
Object detection
YOLO
Model generalization

ABSTRACT

Automated advertisement detection in newspapers is a challenging task due to the diversity in print layouts, formats, and design styles. This task has critical applications in media monitoring, content analysis, and advertising analytics. To address these challenges, we introduce AdVision, a deep-learning-based solution that treats advertisements as unique visual objects. We provide a comparative study of various detection architectures, including one-stage, two-stage, and transformer-based detectors, to identify the most effective approach for detecting advertisements. Our results are validated through extensive experiments conducted under different conditions and metrics. Newspapers from four different countries — Denmark, Norway, Sweden, and the UK — were selected to demonstrate the variety of languages and print formats. Additionally, we conduct a cross-analysis to show how training on one language can generalize to another. To enhance the explainability of our results, we employ GradCAM++ (Chattopadhay et al., 2018) heatmaps. Our experiments demonstrate that the YOLOv8 model achieves superior performance, balancing high precision and recall with minimal inference latency, making it particularly suitable for high-throughput advertisement detection.

1. Introduction

The detection of advertisements in media is a challenging and critical task for content regulation and advertising analytics. In an era where advertisement strategies are rapidly evolving, accurate identification of Ads in newspapers enables businesses, advertisers, and media analysts to track trends, assess Ad effectiveness, and monitor competitors' campaigns. Detecting Ads in newspapers and magazines, whether printed or digital, is complicated because of their high similarity with the rest of the content. Ads variability in size, design, and placement ranges from simple text blurbs to full-page, image-rich layouts. Fig. 1 illustrates examples of diverse advertisement formats and spatial configurations in UK Metro newspaper. Moreover, the line between editorial content and advertisements has become increasingly blurred, as Ads often mimic the style of news articles. This requires advanced detection methods capable of capturing both visual and contextual cues. Language diversity and regional editorial styles further reduce generalizability across datasets.

This task falls within the field of computer vision, where object detection serves as a foundational capability. Despite progress in deep learning for object detection, advertisement detection in printed media remains underexplored, as most existing work targets domains like robotics and healthcare (Nazeri et al., 2024; Xie et al., 2024). Existing works on advertisement detection have primarily focused on media such as television (Rondan et al., 2025; Wardana & Wibowo, 2023), radio (Álvarez et al., 2024), and digital media (Carvalho et al., 2021; Jain et al., 2024), often leveraging structured visual or audio cues like logos, transitions, or speech patterns. In the case of print media, recent efforts have targeted layout-based segmentation or document structuring (Almutairi & Almashan, 2019; Bansal et al., 2021), but these works typically treat advertisements as secondary elements or focus on editorial content. Other approaches, such as ADVISE (Ye & Kovashka, 2018), AdSegNet (Dhang et al., 2024), and LaBINet (Dhang et al., 2025), have explored advertisement understanding and integration in natural scenes, but they assume that Ads are already visually isolated and localized. Similarly, PTPNet (Almgren et al., 2018; Madi et al., 2021) operate on clearly bounded Ad regions in natural or digital images. These assumptions restrict their applicability to the complex,

* Corresponding author at: Mid Sweden University, Holmgatan, Sundsvall, 852 30, Sweden.
E-mail addresses: Faeze.zakaryapoursayyad@miun.se (F.Z. Sayyad), Irida.shallari@miun.se (I. Shallari), Seyedjalaleddin.mousavirad@miun.se (S.J. Mousavirad), Mattias.onils@miun.se (M. O'Nils), faisal.qureshi@ontariotechu.ca (F.Z. Qureshi).

https://doi.org/10.1016/j.mlwa.2025.100686
Received 5 October 2024; Received in revised form 5 June 2025; Accepted 6 June 2025
Available online 18 June 2025

(a) Ad located at the bottom of the page. (b) Full-page Ad. (c) Ad located on the bottom left of the page. (d) Ad located on the bottom right of the page. (e) Without any Ad.

Fig. 1. Examples of UK Metro advertisements displayed in various sizes, locations, and styles. These samples highlight the diversity in placement, ranging from full-page Ads to smaller Ads positioned at different sections of the page. While these examples focus on UK newspapers, there are numerous other styles and features in advertisements across different newspapers and regions.

overlapping, and layout-rich environments found in newspapers, where advertisements often share visual features with surrounding editorial elements and lack consistent structural boundaries. This highlights a gap for investigation with a dataset including a broad range of formats, languages, and styles to capture the variability of advertisements globally. The balanced sampling strategy ensures robustness and mitigates bias, which is crucial for developing reliable and scalable detection models.

Localizing advertisements in a newspaper page involves detecting and possibly classifying the regions containing Ads. For this, robust object detection methods are most appropriate. These methods enable the recognition and localization of objects within the image. Among these uses, one specific application is the detection of advertisements in printed newspapers, which represents a specialized yet significant challenge (Jonsson, 2022). Unlike standard object detection tasks, which often focus on relatively uniform and well-defined object classes, advertisements in newspapers exhibit a high degree of variability in size, design, and embedding context.

As shown in Fig. 1, advertisements can appear in different positions on a newspaper page, such as the bottom of the page, across the whole page, or in specific sections like the bottom-left or bottom-right. These examples from UK newspapers highlight the diversity in size, location, and design of Ads. Additionally, newspaper Ads often vary significantly across publications and regions, exhibiting different visual features and spatial positioning. This variability presents challenges for object detection models, as they must accurately detect Ads under a wide range of conditions.

The progression of deep learning technologies has notably enhanced the capabilities of object detection models, from the early work on R-CNN (Girshick et al., 2014) to the development of more sophisticated and efficient models such as one-stage detectors (e.g., YOLO series Redmon et al., 2016) and transformer-based models (e.g., DETR Carion et al., 2020). Each model presents a different approach to balancing accuracy, computational demand, and adaptability to varied contexts and object categories. These models have shown promise in various object detection tasks, but their application to the task of newspaper Ad detection has been limited. Unlike standard object detection tasks that focus on well-defined object classes, Ads exhibit substantial variability in presentation and embedding context.

Building on prior research (Zakaryapour Sayyad et al., 2024), this study introduces AdVision, to explore the capability of different object detection techniques for newspaper advertisement detection. We first examine different object detection models to find the best architecture for this application based on criteria such as inference latency and detection accuracy. Then, we focus on the best architecture, YOLOv8 (Reis et al., 2023), analyzing it under different conditions. YOLOv8 not only provides the best results for our application but also offers good balance between inference latency and accuracy.

We extend our investigation to include an analysis of the trained Ad detector model's generalizability across different newspapers. Additionally, we introduce a novel exploration of how visual factors, such as color variations and the spatial positioning of advertisements, influence detection accuracy. Contributions of our research include:

1. A comprehensive evaluation of one-stage, two-stage, and transformer-based object detection models for advertisement detection in newspapers.
2. An in-depth analysis of YOLOv8's performance and generalizability when applied to various newspaper formats, providing valuable insights into the practical deployment of advanced object detection models.
3. An exploration of the influence of visual factors, specifically color variations and the spatial positioning of advertisements within the newspaper layout, on the predictive performance of the selected model.

The rest of the paper is structured as follows: Section 2 delves into the state-of-the-art, contextualizing our research within the broader evolution of object detection technologies. Section 3 outlines the methodology, including the criteria for model selection, dataset preparation, and evaluation metrics. Section 4 presents the findings of our experimental evaluations, and Section 5 discusses these results, offering a detailed interpretation of their significance. Finally, Section 6 concludes the paper and suggests avenues for future research.

2. Related work

2.1. Advertisement detection

Advertisement detection has been explored across multiple media formats, each presenting unique challenges and inspiring specialized approaches. Prior research can be broadly categorized by media type: television (Rondan et al., 2025; Siddiqui et al., 2021; Wardana & Wibowo, 2023), radio (Álvarez et al., 2024), newspapers (Barman et al., 2021; Jain et al., 2022), magazines, billboards, and digital platforms (Morera et al., 2020; Yang et al., 2022). Techniques span from video segmentation and speech processing to layout analysis and image recognition.

Numerous approaches to advertisement processing focus on symbolic reasoning, segmentation, or even blockchain-based delivery mechanisms (Liu et al., 2020), yet they often presuppose that advertisements have already been detected. For instance, Ye and Kovashka (2018) introduced ADVISE, a vision-language embedding model that aligns advertisement images with human-written rationale statements to interpret symbolic messages. While influential in advancing semantic understanding, ADVISE assumes advertisements are already localized, limiting its applicability to raw sources like print media. Similarly, Dhang et al. (2024) proposed AdSegNet for detecting billboard corners via segmentation in outdoor video scenes, and later introduced LaBINet (Dhang et al., 2025), which integrates advertisements into natural scenes using Laplace-based Poisson blending. Although LaBINet effectively manages Ad insertion, it also assumes pre-identified Ad regions and does not address the complexities of Ad detection in structured content such as newspapers.

In addition to billboard-focused models, Madi et al. (2021) introduced PTPNet, which uses polygon regression and mask generation

to extract advertisements from natural scene images. All the stated works are developed and evaluated in contexts where advertisements appear as visually distinct, well-bounded regions within structured or uncluttered visual environments. The generalizability of AdSegNet and PTPNet to visually complex or layout-diverse domains remains unexplored. AdVision responds to this gap by offering a solution—automatically detecting and localizing advertisements within visually complex newspaper layouts, thereby enabling frameworks like ADVISE and LaBINet to be applied in broader, real-world contexts.

Machine learning techniques for image classification and object detection have shown potential to improve the reliability and standardization of advertisement recognition (English et al., 2021). Prior studies, such as Almgren et al. (2018) and Jain et al. (2021), used CNN-based models to classify scanned or online newspaper images as Ads or non-ads, but these methods showed limited accuracy and lacked precise localization capabilities. Our work addresses these gaps by employing advanced object detection models that support both classification and spatial localization of advertisements in complex newspaper layouts.

Within the domain of newspaper image analysis, Bansal et al. (2021) introduced a multi-task U-Net framework designed to segment text blocks and detect baselines across diverse layout styles. While effective for structuring editorial content, their approach does not extend to advertisement detection—a task complicated by the visual and structural similarities between Ads and articles. Advertisements in print media often share typographic and layout characteristics with surrounding content, making them difficult to distinguish using models optimized for editorial segmentation. In digital media, Jain et al. (2024) explored advertisement classification by combining visual and textual features with a BERT-based transformer, targeting specific classes of Ads in online newspapers. However, this approach targets predefined Ad classes rather than the full spectrum of advertising content.

Studies in other media formats, such as television (Carvalho et al., 2021; Yang et al., 2024), rely on explicit visual cues like logos or branding elements. However, such cues may be absent or inconsistent in print Ads. Addressing this gap, our work focuses on automatic advertisement localization in newspapers using a general-purpose visual detector that learns ad-specific patterns without relying on predefined markers or structural assumptions.

In the realm of newspaper element segmentation, Almutairi and Almashan (2019) proposed a Mask R-CNN-based framework for segmenting articles, advertisements, and page headers. Their method effectively addressed general newspaper layout analysis across multiple languages and formats. Building on this, Almutairi (2024) introduced a more advanced framework utilizing CNNs and Transformers to not only detect elements but also categorize pages into sections. While highly relevant to our study, this approach treats advertisements as secondary elements and does not address the specific challenges of advertisement detection across heterogeneous newspaper styles. This leaves a gap in developing a robust, efficient system specifically for Ad localization and generalization. Our work, AdVision, addresses this by focusing exclusively on advertisement detection, leveraging state-of-the-art object detection architectures, and validating them across multiple newspaper datasets and layouts. Furthermore, we introduce model explainability techniques to interpret complex visual content, setting a foundation for downstream tasks like Ad interpretation or content filtering.

Recent studies have demonstrated the potential of deep learning for detecting advertisements in newspaper layouts. Barman et al. (2021) explored automated document processing by analyzing historical newspaper data through combined text and image segmentation techniques. Their research discusses the challenges of segmenting and classifying newspaper pages. However, relying solely on historical data can introduce temporal bias, which affects the applicability of these methods to current scenarios due to rapid changes in editorial practices and visual presentation styles. In contrast, our study builds on prior work in newspaper layout analysis by using datasets that reflect modern newspaper designs, resulting in a more practical and generalizable solution.

Our research advances the field of advertisement detection within newspapers by addressing several critical gaps identified in the literature. Unlike previous methods, our approach integrates advanced computer vision techniques to create a comprehensive and accurate detection system. By developing and utilizing extensive, varied datasets that reflect the complexities of real-world advertising scenarios, our research enhances the reliability and applicability of Ad detection in print media, setting a new standard for future studies in this domain. To the best of our knowledge, no prior work has systematically evaluated the generalizability of advertisement detection models across diverse newspaper datasets, including unseen newspapers from different languages and formats. By documenting the strengths and limitations of YOLOv8 in these scenarios, we fill a critical gap in the literature and offer actionable insights for improving cross-dataset adaptability.

2.2. Object detection

Object detection in computer vision has seen significant advancements, driven by the development of deep learning technologies. While traditional CNN-based object detectors continue to dominate the field, emerging models such as capsule-enhanced networks have demonstrated improved robustness in sparse and noisy environments (Zhang et al., 2024), suggesting their potential applicability to visually complex layouts such as printed media. Research in the domain of deep learning-based object detection can be broadly classified into two categories: one- and two-stage object detection algorithms (Zou et al., 2023). Recently, with the introduction of transformer-based methodologies in computer vision, a new category of object detectors, known as transformer-based object detectors, has emerged, which can be integrated into both one-stage and two-stage frameworks.

2.2.1. One and two stage detectors

Two-stage algorithms have emerged as a structured approach for accurate object localization in image-based object detection. Notable examples include Grid R-CNN (Lu et al., 2019) and Faster R-CNN (Ren et al., 2015). Faster R-CNN (Ren et al., 2015) is a key model within the two-stage paradigm. It integrates a Region Proposal Network (RPN) to efficiently generate potential object regions. The subsequent stage refines these regions through regression and classification. This unified architecture eliminates the need for separate region proposal methods, improving efficiency and accuracy.

Grid R-CNN introduces an innovative grid-guided localization mechanism within the two-stage framework. It explicitly captures spatial information and leverages the position-sensitive attributes of fully convolutional architectures. By predicting grid points and using them to guide bounding box determination, Grid R-CNN achieves high-quality object localization. A multi-point supervision approach mitigates inaccuracies, and a two-stage information fusion strategy enhances performance. Grid R-CNN's effectiveness is demonstrated through benchmark results and its adaptability to various detection frameworks.

One-stage detectors such as You Only Look Once (YOLO) (Redmon et al., 2016) revolutionized object detection by combining the region proposal and classification stages into a single network. This approach significantly improved detection speed while maintaining competitive accuracy. Variants of YOLO, particularly YOLOv8 (Reis et al., 2023), have shown impressive performance in various object detection tasks due to enhancements in network architecture and training methodologies. RetinaNet (Lin et al., 2017) introduced a focal loss function to address class imbalance in object detection, providing another robust one-stage detector that balances speed and accuracy. RTMDet (Lyu et al., 2022) and other similar models further optimized this balance, making them suitable for diverse applications, including real-time scenarios.

2.2.2. Transformer-based detectors

Transformers have recently drawn more interest in the vision and object detection task. As a result, a brand-new class called transformer-based object detection techniques has evolved. Transformers are one type of deep learning architecture. It was first proposed in "Attention is all you need" paper (Vaswani et al., 2017) to solve NLP tasks. DETR (Carion et al., 2020) is the first fully end-to-end transformer-based object detector (Bar et al., 2022). It discards several hand-designed components like Spatial Anchor and builds a simpler pipeline for the detection task. DETR combines convolutional neural networks (CNN) and Transformer-Encoder-Decoders. Building on DETR, DAB-DETR (Liu et al., 2022) combines a CNN backbone with Transformer encoders to extract and refine spatial features. It uses dual queries (positional and content) to interact with the decoder, progressively refining these queries to generate object predictions.

The Swin Transformer further enhances transformer-based object detection by providing a general-purpose backbone for computer vision tasks. Its hierarchical architecture and shifted windows achieve high performance on benchmarks like the COCO dataset (Lin et al., 2014), maintaining computational complexity that is linear in input size. In our experiments, we used DAB-DETR and RetinaSwin, employing the Swin Transformer (Liu et al., 2021) as a backbone for RetinaNet (Lin et al., 2017). RetinaNet combines the Feature Pyramid Network (FPN) with ResNet backbones and specialized sub-networks for classification and bounding box regression, leveraging the Swin Transformer for improved performance.

While the presented models demonstrate high accuracy in object detection, their adaptability to the unique characteristics of newspaper layouts remains unclear. Additionally, most studies lack emphasis on inference efficiency, a critical factor for high-throughput applications. The need for a robust and efficient detection framework capable of handling the complexities of newspaper advertisements motivates this study. Understanding the strengths and limitations of different object detection paradigms in this context is crucial for advancing the field.

3. Methodology

This paper introduces AdVision, a methodology for automatically finding the advertisement in printed media. Our objective is to provide the performance trade-off between different state-of-the-art object detection architectures to detect advertisements within newspaper publications. We provide a cross-analysis on these architectures, and also present an analysis based on GradCAM++ for the explanability of the results. To achieve this, we implemented a comprehensive methodology comprising dataset collection and annotation, followed by the application of object detection models. This process enabled us to accurately identify and extract the coordinates of advertisements, allowing for precise localization within the newspaper pages.

3.1. Dataset collection and annotation

The dataset for this study is a diverse collection of newspapers from multiple countries, showcasing a variety of languages and print formats. The specific newspapers included in the dataset are: (1) Berlingske (Berlingske, 2022–2023) from Denmark, (2) Adresseavisen (Adresseavisen, 2022–2023) from Norway, (3) Sydsvenskan (Sydsvenskan, 2022–2023) from Sweden, and (4) UK Metro (Metro, 2022–2023) from England. The primary objective of compiling this multi-source dataset is to introduce linguistic diversity and capture the different editorial and advertising styles found across countries. Each newspaper is published in the official language of its respective country and features unique formatting styles, allowing for the analysis of varied advertisement designs and text layouts. To prevent a bias in the dataset, it includes balanced samples from each newspaper on different days of the week over approximately one year. This approach minimizes bias and enhances the comprehensiveness of the analysis.

Table 1
Dataset statistics containing number of nulls and number of not nulls images for the whole annotated data before sampling.

Dataset name	N_{null}	N_{adv}	N_{Total}	P_{null}	Dataset period
Berlingske (X_1)	2975	1785	4760	62.5%	04/2022–03/2023
Adresseavisen (X_2)	2676	2573	5249	50%	04/2022–03/2023
Sydsvenskan (X_3)	2117	1396	3513	60%	10/2022–01/2023
UK Metro (X_4)	672	1943	2615	34%	04/2022–03/2023
Aftonbladet	419	145	564	74.3%	05/2023
iTromsø	476	147	623	76.4%	05/2023

The datasets from the individual newspapers are represented as $X = \{X_1, X_2, X_3, X_4\}$, where X_i denotes the set of images from the ith newspaper. For each newspaper X_i, the data is divided into:

- **Null images** (N_i^{null}) Images of newspaper pages without advertisements; and
- **Advertised images** (N_i^{adv}) Images containing one or more advertisements.

The total number of images in each dataset, N_i^{total}, is the sum of null and advertised images:

$$N_i^{total} = N_i^{null} + N_i^{adv}. \tag{1}$$

The proportion of null images in each newspaper dataset, represented as P_i^{null}, is calculated using the formula:

$$P_i^{null} = \frac{N_i^{null}}{N_i^{total}} \times 100. \tag{2}$$

This metric, P_i^{null}, quantifies the percentage of pages lacking advertisements, which is a critical factor in understanding advertisement frequency and placement strategies across different markets.

A comprehensive overview of the collected data and their respective timespans is presented in Table 1. Each page within the newspapers is digitized as an image. Additionally, these images have been annotated with bounding boxes to delineate each advertisement, facilitating detailed analysis of advertising content and layout across diverse newspaper styles. To further test and check the generalizability of our methodology, we downloaded and annotated additional newspapers from Aftonbladet (2023) from Sweden and iTromsø (2023) from Norway for the dates 25 to 31 May 2023. These additional datasets were used solely for testing purposes to ensure the robustness of our object detection models across different sources and formats.

3.2. Data sampling and balancing

Our initial focus in this study was on the Adresseavisen newspaper (X_2), chosen for its balanced mix of advertised (adv) and non-advertised (null) content. However, we encountered challenges with the UK Metro dataset (X_4) where only 34% of pages were null, indicating a significant imbalance that could potentially skew the analysis. To address this issue and ensure a more balanced dataset across different newspaper types, we implemented the following strategies:

- **Additional Sampling of Null Pages:** We supplemented the UK Metro dataset by randomly adding 800 null pages from additional dates within the reported period. This adjustment was crucial for normalizing the proportion of null to advertised pages, making it comparable to the other datasets in our study.
- **Equal Sampling Across Categories:** For each newspaper dataset (X_1 to X_4), we sampled an equal number of advertised and non-advertised pages, specifically 1396 of each type. This strategy helps maintain data harmony and prevents any model bias towards a particular category.

A random seed (value of 42) was used to ensure that our sampling process was reproducible and that the image files were shuffled consistently across different data collection iterations. This method helps confirm that our findings are reflective of genuine dataset features and not merely artifacts of the selection process.

After addressing these sampling and balancing issues, we divided the images into distinct subsets for training, validation, and testing, using a 70%–20%–10% split ratio. This distribution strategy was applied to both the advertised and non-advertised pages to ensure that each subset was equally represented. By maintaining this balanced approach throughout our study, we aimed to enhance the validity and reliability of our experimental findings. This is essential for avoiding potential biases and providing a thorough understanding of the model's performance across different types of newspaper content. Our approach could be extended to health-related advertising or regulatory applications. For instance, studies leveraging statistical feature analysis in behavioral data (Zhao et al., 2025) provide a foundation for understanding how model outputs might link with real-world physiological or behavioral phenomena. Future extensions may consider adversarial techniques such as DiGAN, which demonstrated effective correction of data imbalance using GANs in biomedical datasets (Zhao et al., 2024).

3.3. Data preprocessing

Following the results outlined in our previous research (Zakaryapour Sayyad et al., 2024), we applied similar preprocessing techniques in this analysis. To ensure consistency, all newspapers in the dataset were scaled to the same resolution and aspect ratio. We used bilinear interpolation to adjust the images to a resolution of 608×608 pixels. Additionally, white padding was applied to maintain a 1:1 aspect ratio, preventing distortions during the training and testing phases.

3.4. Model selection

The selection of appropriate models for advertisement detection is a critical step in our methodology. We aim to identify models capable of accurately detecting and localizing advertisements within newspaper pages. Our approach involved evaluating several state-of-the-art object detection algorithms to determine the most effective one for our specific dataset and objectives.

Several factors were considered in the selection process, including model accuracy on benchmark datasets, computational efficiency, the model's approach to object detection, and its applicability to the problem at hand. The evaluated models include:

1. **Faster R-CNN** (Ren et al., 2015)**:** Known for its accuracy and robustness, Faster R-CNN is a region-based convolutional neural network that excels in object detection tasks. Its two-stage approach—first generating region proposals and then classifying these regions—makes it a strong candidate for detecting advertisements. This model offers high accuracy in various object detection benchmarks but at the cost of increased computational demand.
2. **Grid R-CNN** (Lu et al., 2019): Similar to Faster R-CNN, Grid R-CNN uses a two-stage architecture with region proposal networks and feature extraction. It refines object locations by predicting the positions of grid points, which can improve localization accuracy, especially for complex shapes.
3. **YOLOv8** (Reis et al., 2023): The latest iteration in the YOLO series, YOLOv8 builds on the speed and accuracy of its predecessors, incorporating improvements in model architecture and training techniques to enhance performance, particularly for real-time object detection tasks. The summarized architecture of YOLOv8 is shown in Fig. 2. One of the key features of YOLOv8 is the Anchor-Free Detection. Traditional object detectors use anchor boxes of various shapes and sizes to predict object locations. Anchor-free methods, however, directly predict the center of objects, which simplifies the model and reduces computation during post-processing. This approach can enhance the model's speed and accuracy by eliminating the need to match anchor boxes to ground truth boxes during training. Another key feature of the YOLOv8 model is that this model starts with smaller initial convolution layer which helps in better capturing fine-grained features early in the network.
   YOLOv8 backbone structure consists of several convolutional layers and the new C2f modules. The smaller initial convolution helps in capturing initial features, followed by more complex layers for deeper feature extraction. C2f Module is an improved version of the CSP (Cross-Stage Partial) module, designed to enhance feature extraction efficiency and effectiveness. Residual and CSP Blocks improve gradient flow and reduce computational costs, making the model more efficient and deeper without performance degradation. The third important feature of YOLOv8 is Mosaic Augmentation. Mosaic augmentation is a data augmentation technique that combines four images into one during training. This helps the model learn to detect objects at different scales and contexts. By exposing the model to varied image compositions, it becomes more robust. Turning off this augmentation in the last epochs of training helps fine-tune the model on more natural images, improving final performance.
4. **RTMDet** (An Empirical Study of Designing Real-Time Object Detectors) (Lyu et al., 2022): RTMDet is designed for high performance in real-time object detection. It introduces an efficient architecture with large-kernel depth-wise convolutions and dynamic label assignment with soft labels, achieving superior real-time object detection performance and versatility across various recognition tasks. It showcased its superiority over industrial detectors by demonstrating a parameter-accuracy trade-off across various model sizes, making it versatile for diverse application scenarios.
5. **DAB-DETR** (Dynamic Anchor Boxes are better queries for DETR) (Liu et al., 2022): DAB-DETR introduces a query formulation using dynamic anchor boxes within the DETR framework. This method improves detection accuracy and efficiency by utilizing box coordinates as queries that are dynamically updated layer-by-layer in Transformer decoders, offering better handling of objects of varying sizes and aspect ratios.
6. **DDOD** (Disentangle your dense object detector) (Chen et al., 2021): DDOD focuses on disentangling classification and localization tasks to improve detection accuracy. It uses dynamic convolutions to adaptively adjust to the object's appearance, leading to significant performance improvements on benchmarks like MS COCO (Lin et al., 2014).
7. **RetinaNetSwin** (Lin et al., 2017; Liu et al., 2021): Combining the strengths of RetinaNet and the Swin Transformer, this model significantly enhances object detection capabilities. RetinaNet introduces a focal loss function to address class imbalance, improving both detection accuracy and robustness, making it highly effective for identifying objects in dense and complex scenes. Meanwhile, the Swin Transformer boosts object detection through its hierarchical vision transformer backbone, employing shifted windows to enhance feature extraction and scalability. Together, these innovations result in a model that captures multi-scale features more effectively, pushing the boundaries of object detection performance.

Based on a comprehensive set of experiments, we aim to select the robust and reliable model for Ad detection in terms of different criteria.

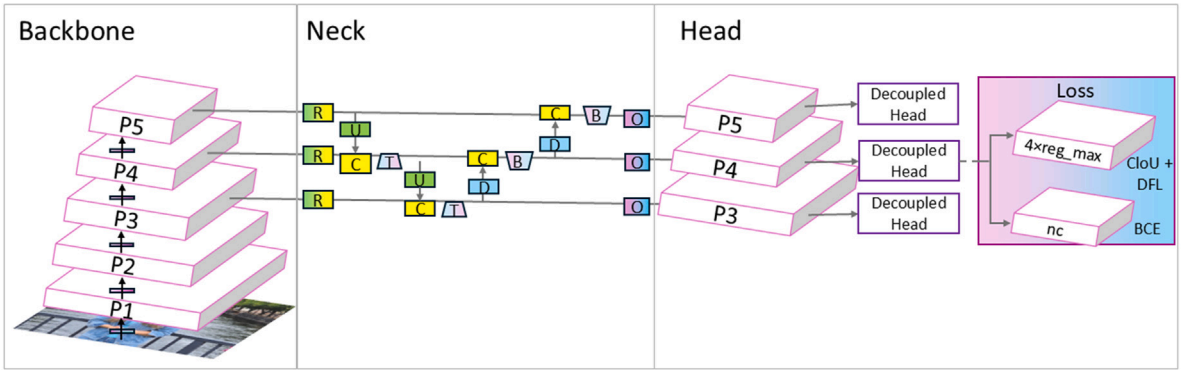


Fig. 2. An overview of YOLOv8 architecture.

Table 2
Model training setup.

Model	Batch size	Backbone	Base LR	Optimizer
Faster R-CNN	8	ResNeXt101 (Xie et al., 2017)	8e−05	AdamW
Grid R-CNN	8	ResNeXt101 (Xie et al., 2017)	8e−05	AdamW
DDOD	24	ResNet50 (Wightman et al., 2021)	2e−02	SGD
RetinaNetSwin	8	Swin (Liu et al., 2021)	8e−05	AdamW
DAB-DETR:	2	ResNet50 (Wightman et al., 2021)	1e−05	AdamW
RTMDet	4	CSPNeXt (Wang et al., 2020)	4e−05	AdamW
YOLOv8	16	CSPDarknet53 (Bochkovskiy et al., 2020)	Auto	Auto

3.5. Training and validation

The analysis presented in this study focuses on two main aspects: first, the evaluation of various advertisement detection models under different conditions, specifically examining both detection accuracy and inference time; and second, the assessment of selected model's generalization capabilities across diverse datasets. Initially, we identified the best-performing model using a single newspaper, Adresseavisen. Subsequently, we tested the generalization of the selected model by evaluating its detection accuracy on other types of newspapers. To ensure unbiased results, we employed K-fold validation in our experiments.

We trained each model on our preprocessed dataset, using the training subset for model training and the validation subset to tune hyperparameters and prevent overfitting. The test subset was reserved for final evaluation. We ensured that the training and validation processes maintained the balanced representation of advertised and non-advertised pages to avoid bias.

For configuring the models, our goal was to replicate the optimal parameters described in their respective publications. However, due to limitations in computational resources, adjustments such as reducing the batch size were necessary, as detailed in Table 2. The ROI head of two-stage models and the bounding box head of single-stage models were specifically designed for a single class (advertised) and initialized with pretrained COCO (Lin et al., 2014) weights. All models were trained for 150 epochs.

3.6. Evaluation metrics

This research evaluates the performance of various advertisement detection models under different conditions. To this end, we utilized three primary metrics: Mean Average Precision (mAP), F1 score, and inference latency. These metrics were chosen to achieve a balance between model accuracy and computational efficiency. By considering these metrics we aim to develop a model that is not only accurate but also efficient and practical for large-scale deployment.

Mean Average Precision (mAP) is a performance metric that provides an aggregate measure of the model's Precision-Recall trade-off across different object detection thresholds. It is calculated as follows:

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i, \tag{3}$$

where n is the number of classes or categories, and AP_i denotes the Average Precision for each class. This metric allows us to evaluate the model's performance across all categories comprehensively.

$F1$ score is another metric that we employed during the evaluation. It serves as a composite measure that balances Precision (P) and Recall (R), and is calculated as:

$$F1 = \frac{2PR}{P+R}, \tag{4}$$

where Precision is the ratio of correctly predicted positive observations to the total predicted positives, and Recall is the ratio of correctly predicted positive observations to all observations in the actual class. The $F1$ score provides a single measure that accounts for both false positives and false negatives.

Inference latency is also a crucial metric in our scenario. Given that the final solution is intended to process massive amounts of newspaper material, achieving low latency is essential for high throughput and faster response times. This metric ensures that the model can operate efficiently in real-world applications, where quick processing times are necessary.

4. Results

Our evaluation explored the performance of several advertisement detection models across different conditions. The primary findings are summarized in Table 3, which presents a comparative analysis of these models. According to the results reported in Table 3, YOLOv8 outperforms all other object detectors in the application of advertisement detection. In terms of overall detection accuracy, the YOLOv8 model outperformed all other models with the highest precision (0.907) and mAP50 of (0.956) on Addresseavisen test set, while also achieving the fastest inference speed at 1.5 ms. This indicates a superior balance of accuracy and efficiency, making it highly suitable for high-throughput advertisement detection applications. In contrast, models such as Faster R-CNN and Grid R-CNN, despite their robust performance metrics (precision of 0.87 and 0.83, respectively), showed comparatively slower inference times.

To further validate the superiority of YOLOv8, we statistically compared YOLOv8 using two different tests: a paired t-test for pairwise comparisons and Friedman for multiple model comparisons. The t-test is a statistical test used to compare the means of two groups and determine if the differences between them are statistically significant. It calculates the t-statistic, which measures the size of the difference

relative to the variation in the data. A larger t-statistic indicates a more significant difference between the two groups.

In this problem, a paired t-test was conducted to compare the performance of YOLOv8 with other models across multiple data splits. The null hypothesis here states that there is no significant difference between YOLOv8 and the other models. A p-value lower than 0.05 indicates that we can reject the null hypothesis, suggesting a significant difference between the models. From Table 4, we can observe that all the p-values are far below the threshold of 0.05, confirming that the performance differences between YOLOv8 and each of these models are statistically significant. Also, the degrees of freedom (df) for this problem is 4 and the critical value for a two-tailed test at $\alpha = 0.05$ and df = 4 is 2.776 (from the t-distribution table) . From Table 4, we can see that the calculated t-statistics are significantly larger than the critical value, further supporting the conclusion that YOLOv8's performance is significantly different from (and better than) the other models.

Another statistical test used to evaluate the effectiveness of the YOLOv8 algorithm is the Friedman test. As previously stated, this is a multiple comparison test. Table 5 shows the average ranks calculated using the Friedman statistical test. According to this table, the YOLOv8 algorithm achieves the highest rank. Additionally, the *P*-value in Table 5 suggests that the differences in rankings are statistically significant at a 5% significance level.

Tables 6 to 9 present our analysis on generalization capabilities of models across diverse datasets. These tables explore the performance of models on various datasets, illustrating significant variability in adaptability. We evaluate the model based on seen and unseen test sets. Seen test sets are pages that come from the same newspaper title(s) that were used during training, whereas unseen test sets correspond to pages from newspaper titles that were completely absent from the training split. For Tables 6 to 9 each model was trained on a single newspaper (Berlingske, Adresseavisen, Sydsvenskan, or UK Metro) and evaluated on: the seen portion of that same title, and the unseen titles (the remaining three newspapers). Our goal is to measure out-of-the-box generalization; introducing even a small amount of target data would have confounded that comparison. No post-hoc adaptation (e.g., few-shot fine-tuning, test-time updating, or feature-alignment methods) was applied; investigating such strategies to close the seen–unseen gap is left for future work.

Models trained and tested on the same dataset generally performed better. For instance, Sydsvenskan model achieved a precision of 0.96, a recall of 0.97, and an mAP50 of 0.98. However, when tested on different datasets, performance declined sharply, such as the model trained on the Berlingske dataset and tested on the Adresseavisen dataset, which only achieved a precision of 0.43 and a recall of 0.62. This variability underscores the challenges of dataset diversity in model robustness.

It is worth mentioning that the performance degradation was not uniform across all model-dataset combinations, indicating that certain datasets have inherent qualities that may better handle the diversity of unseen data. For example, when the Sydsvenskan model was tested on the UK Metro dataset, it maintained a precision of 0.75 and a recall of 0.84, suggesting some level of adaptability.

Building on this observation, we extended our investigation to assess the model's generalization capabilities across a broader range of datasets, including both seen and unseen newspapers. Here, 'seen' refers to the newspaper frames included during training, while 'unseen' refers to new newspapers that the model has not encountered during training, featuring different styles and characteristics. The objective was to explore how training data diversity influences the development of a robust advertisement detection model. The distinction between seen and unseen test sets, based on whether they consist of similar or different newspapers from those used in training, illuminates the challenges and opportunities in generalizing AI model to varied visual content. A robust model must accurately detect advertisements in a wide range of scenarios, including those that differ from the training data.

When evaluated across all samples, the YOLOv8 model achieves a good performance, with precision and recall almost at par (0.937 and 0.933, respectively). This high level of accuracy demonstrates the model's ability to correctly identify advertisements with minimal false positives or negatives. The mAP50 and mAP50-95 scores, both exceeding 0.93, further confirm the model's robust detection capabilities across a broad range of scenarios.

Table 3
Comparison of advertisement detectors on adresseavisen newspaper dataset for best model selection.

Model	Precision	Recall	mAP50	Inference time (ms)
Faster R-CNN (Ren et al., 2015)	0.87	0.86	0.90	47
Grid R-CNN (Lu et al., 2019)	0.83	0.83	0.88	47.5
DDOD (Chen et al., 2021)	0.85	0.82	0.88	26.2
RetinaNetSwin (Lin et al., 2017; Liu et al., 2021)	0.85	0.81	0.90	32
DAB-DETR (Liu et al., 2022)	0.84	0.86	0.87	30.6
RTMDet (Lyu et al., 2022)	0.84	**0.91**	0.88	43.3
YOLOv8 (Reis et al., 2023)	**0.91**	0.87	**0.96**	**1.5**

Table 4
T-test comparison of YOLOv8 and other models in terms of mAP@50.

Model	*t*-statistic	p-value
RetinaSwin	6.370749	0.003113
dab	14.086510	0.000147
ddod	4.692708	0.009359
faster_rcnn	4.041092	0.015590
grid	6.536351	0.002831
rtmdet	5.412984	0.005643

Table 5
The Friedman ranks in terms of mAP50 metric.

Models	Rank
YOLOv8	1.0
grid	5.4
dab	5.4
faster_rcnn	3.8
ddod	4.7
RetinaSwin	3.5
rtmdet	4.2
P-value	0.0212

Table 6
Performance of the YOLOv8 model trained on the Berlingske dataset and tested on other newspaper datasets.

Test dataset	Precision	Recall	mAP50	mAP50-95
Berlingske (X_1)	0.89	0.97	0.97	0.94
Adresseavisen (X_2)	0.43	0.62	0.49	0.39
Sydsvenskan (X_3)	0.56	0.60	0.58	0.49
UK Metro (X_4)	0.66	0.49	0.56	0.48

Table 7
Performance of the YOLOv8 model trained on the Adresseavisen dataset and tested on other newspaper datasets.

Test dataset	Precision	Recall	mAP50	mAP50-95
Berlingske (X_1)	0.60	0.71	0.63	0.56
Adresseavisen (X_2)	0.89	0.85	0.91	0.87
Sydsvenskan (X_3)	0.80	0.76	0.85	0.76
UK Metro (X_4)	0.57	0.66	0.61	0.54

5. Discussion

Our experimental findings, presented in the earlier section, suggests that the YOLOv8 model might be learning specific visual cues from

Table 8
Performance of the YOLOv8 model trained on the Sydsvenskan dataset and tested on other newspaper datasets.

Test dataset	Precision	Recall	mAP50	mAP50-95
Berlingske (X_1)	0.73	0.67	0.71	0.64
Adresseavisen (X_2)	0.68	0.79	0.78	0.64
Sydsvenskan (X_3)	0.96	0.97	0.98	0.95
UK Metro (X_4)	0.75	0.84	0.85	0.80

Table 9
Performance of the YOLOv8 model trained on the UK Metro dataset and tested on other newspaper datasets.

Test dataset	Precision	Recall	mAP50	mAP50-95
Berlingske (X_1)	0.49	0.62	0.61	0.55
Adresseavisen (X_2)	0.64	0.57	0.57	0.44
Sydsvenskan (X_3)	0.80	0.67	0.78	0.70
UK Metro (X_4)	0.97	0.98	0.99	0.97

Table 10
Test results for the YOLOv8 model trained on the Berlingske, Adresseavisen, Sydsvenskan, and UK Metro (X) train set. The model's performance is evaluated on multiple datasets, including both the newspapers used in training (seen) and entirely new newspapers from a different publisher (unseen).

Test dataset	Precision	Recall	mAP50	mAP50-95
All samples (X)	0.94	0.93	0.97	0.94
Berlingske (X_1)	0.91	0.94	0.97	0.94
Adresseavisen (X_2)	0.89	0.89	0.93	0.88
Sydsvenskan (X_3)	0.96	0.98	0.98	0.95
UK Metro (X_4)	0.99	0.98	0.99	0.98
Aftonbladet (unseen)	0.84	0.80	0.87	0.82
iTromsø(unseen)	0.79	0.69	0.78	0.66

the training data, limiting its ability to generalize to unseen layouts or styles. These insights are crucial in interpreting the results documented in Table 10, where the performance of the model on both seen and unseen test sets is evaluated. The high performance on seen test sets, including Berlingske, Adresseavisen, Sydsvenskan, and UK Metro, with precision and recall rates above 0.89 in all cases, demonstrates the model's proficiency in accurately identifying advertisements in familiar contexts. These results, particularly the near-perfect metrics on the UK Metro dataset, attest to the model's effectiveness in environments closely aligned with its training data. Such environments, characterized by similar visual distributions and content styles, allow the model to leverage learned patterns effectively, resulting in high precision and recall.

Conversely, the model's performance suffers when encountering unseen data, as evidenced by the drop in accuracy on Aftonbladet and iTromsødatasets. These datasets represent newspapers with distinct visual styles not included in the training data. This highlights a limitation: the model's learned patterns struggle with novel visual distributions, leading to reduced precision and recall. To address this and create more adaptable models, future research should focus on incorporating a broader range of visual styles and content types in training data. The observed performance difference between seen and unseen datasets offers valuable insights for improving advertisement detection models. Strategies such as expanding the diversity of training datasets, employing domain adaptation techniques, and leveraging unsupervised learning to better understand and adapt to novel visual distributions are critical.

Our experimental findings provide additional layers of understanding regarding the model's behavior under varied operational conditions, particularly in terms of feature recognition and model interpretability. Initially, we applied the Grad-CAM++ (Chattopadhay et al., 2018) technique to elucidate which features the YOLOv8 model prioritizes during the Ad detection process. Grad-CAM++ is a visualization technique designed to enhance the interpretability of convolutional neural networks by highlighting the regions of an input image that are most influential for predictions.

These heatmaps highlight areas that a model might consider important when analyzing the content. The red and yellow regions indicate high attention or importance, while blue areas denote less attention. Notably, the heatmaps show attention around the central figures or key text areas in advertisements and headlines. For example, key text elements such as prices, product names, and promotional messages tend to have high attention scores. Visual elements like faces, bright colors, and central figures also attract significant attention.

However, the results indicate some ambiguity in the model's feature prioritization. The heatmaps did not consistently differentiate between advertisements and other content, complicating our understanding of the model's focal points. The generated heatmaps do not distinctly separate Ads from adjacent content. This is evident in the provided samples, where the attention is diffused and does not offer clear guidance on the areas of highest importance, as seen in the images in Fig. 4. For example, in the Adresseavisen samples, high activations cluster around large numeric price tags and saturated product imagery—even spilling over into adjacent editorial graphics with similar color profiles—indicating a bias toward bold chromatic edges rather than true Ad boundaries. Conversely, the more regimented, single-column Sydsvenskan inserts produce a tighter, column-shaped attention pattern, suggesting that clear geometric layout cues anchor the model more reliably than semantic content alone.

To further explore this issue, we conducted a series of experiments by manipulating the advertisement placements and contents within the newspaper pages as shown in Fig. 3. One notable experiment involved altering the position of an advertisement on a weather prediction page from the bottom to another location. Interestingly, the model failed to detect the advertisement post-manipulation. This suggests a potential overfitting to the advertisement's original position or a reliance on certain contextual cues surrounding the original Ad placement.

Subsequently, replacing the original advertisement with another from a different newspaper resulted in successful detection by the model. This indicates that the model is capable of recognizing advertisements across different contexts when they conform to previously learned features typical of Ads. However, substituting the advertisement with a blank white space led the model to incorrectly classify the area as containing an advertisement. This misclassification could be attributed to the model's reliance on certain surrounding elements and features that are still present above the white space. These cues, such as layout elements (e.g., borders, text formatting) or surrounding editorial content (e.g., headlines, captions), might have been used by the model to identify the original advertisement, even if the advertisement itself was not present. When replaced with blank space, these contextual cues might have misled the model, leading to a false positive detection.

This suggests that the model might not be solely focusing on the content of the advertisement itself, but also on the surrounding visual context. While this can be helpful in some cases, it also highlights a potential limitation. Dependence on contextual cues can lead to misclassifications when those cues are present without an actual advertisement. Future research could explore methods to make the model less reliant on contextual cues and focus more on the specific visual characteristics of advertisements themselves. This could involve techniques like training the model on datasets with a wider variety of layouts and backgrounds, or incorporating mechanisms that explicitly separate content recognition from contextual analysis.

Furthermore, replacing the advertisement with editorial content did not trigger any detection, suggesting that the model can differentiate between textual content layouts typical of advertisements and those of editorials. This differentiation likely arises from distinct features learned by the model, such as text arrangement, font size, and accompanying graphical elements, which are different in Ads compared to editorial content. These observations indicate that while the model is quite robust in detecting advertisements in their typical contexts, it struggles with Ads that are re-positioned or significantly altered in content. This suggests a potential area for improvement in the model's

training to better generalize across varied layouts and content types. To address this limitation, we must consider that future Ad placements within newspapers by editors are unpredictable. Therefore, it might be necessary to augment the dataset with advertisements appearing in parts of the editorial content where they have not previously or rarely appeared. To further explore the dimensions of model generalizability and robustness, we extended our investigation to examine how variations in color influence the YOLOv8 model's ability to detect advertisements. This inquiry was motivated by the hypothesis that color, as a distinctive feature in advertisements, might play a critical role in the model's detection capabilities (see Fig. 5 for samples of the altered colors).

Table 11 presents the ten-fold cross-validation results of the model trained on the Adresseavisen dataset using different image color formats. The table includes the performance metrics alongside the variance values. The variance is calculated as the deviation between the original RGB images and the altered images, providing an indication of how much the model's performance is affected by these color transformations. For two RGB images, the variance is calculated by flattening the images into 1D arrays, computing the difference between corresponding pixels, and then finding the variance of these differences. For an RGB image and a grayscale image, the variance is computed for each RGB channel and the grayscale image separately, and the average of the absolute differences between these variances is calculated.

From Table 11, it is evident that the model achieves the highest performance metrics when processing images in the original RGB format, with Precision, Recall, mAP50, and mAP50-95 values all in the high nineties. The metrics show a progressive decrease in performance as the image representation moves further from the original RGB format, which can be interpreted in terms of how the model utilizes color information. The minimal drop in performance when testing on grayscale images suggests that while the model does utilize color information, it is also effectively leveraging luminance (brightness) information. The variance value of 232.7, although it indicates some level of deviation from the RGB baseline, still shows that the model retains much of its effectiveness, implying that the structure and luminance captured in grayscale are sufficient for most detections.

The binary images (converted through thresholding which reduces the image to pure black and white based on a luminance threshold) show a more pronounced drop in all performance metrics. The higher variance value of 4265.4 indicates a significant deviation from the performance on RGB images. This suggests that the removal of intermediate shades of gray (which include detailed luminance information) impacts the model's ability to detect and classify objects accurately, highlighting the importance of more nuanced luminance details beyond mere color.

However, the most notable decline in performance is observed when the model processes images with inverted colors. In this case, color inversion is achieved by subtracting each pixel's RGB value from 255, reversing color intensities (e.g., black becomes white and vice versa). Under these conditions, precision falls to 0.74, and Recall sees a significant reduction to 0.58. Furthermore, the mAP50 and mAP50-95 are drastically lower, with mAP50 at 0.67 and mAP50-95 at 0.26. The very high variance of 7601.2 and poorer performance metrics with inverted color images underline the importance of specific color hues and their standard luminance patterns in the model's training. This suggests that the model is highly dependent on the natural appearance of colors and their arrangements for accurate detection, and color inversion disrupts these learned patterns.

The results indicate that there is an inverse correlation between variance and performance criteria: as the variance increases, indicating greater deviation from the RGB baseline, the performance metrics tend to decrease. This inverse correlation highlights that higher variance corresponds to conditions under which the model's effectiveness is compromised, underscoring the model's reliance on familiar color and luminance patterns for optimal performance. These findings imply that while the YOLOv8 model can handle the absence of color to some extent, it is dependent on the original color patterns for accurate advertisement detection. Future studies could explore mechanisms to enhance the model's robustness to color variations, such as training the model on a more diverse dataset that includes multiple color transformations or improving the model's architecture to be less sensitive to color changes.

Table 11
Effect of color on model performance: Ten-fold cross-validation results for the YOLOv8 model trained on Adresseavisen using original RGB images.

Dataset	Precision	Recall	mAP50	mAP50-95	Variance to RGB
Original RGB	**0.96**	**0.95**	**0.98**	**0.96**	0
Grayscale	0.94	0.94	0.98	0.94	232.7
BinaryColor	0.87	0.78	0.87	0.79	4265.4
Inverted color	0.74	0.58	0.67	0.26	7601.2

6. Conclusion

In this study, we evaluated the effectiveness of various state-of-the-art object detection models for detecting advertisements in printed media. The results indicate that YOLOv8 outperformed other models in terms of precision, recall, and inference speed, making it highly suitable for high-throughput advertisement detection applications. Tests across different newspaper datasets revealed that while the model adapts well to newspapers within the source domain, its performance declines on unseen target domains. This suggests that incorporating all target domain newspapers into the training set may enhance the model's ability to detect Ads across a broader range of newspapers. From the analysis, it was observed that the model rely on specific features or colors. Future research should focus on creating a general model that can effectively detect advertisements across diverse newspaper formats. This can be achieved by including a broader range of training data, exploring domain adaptation techniques, and leveraging advancements in synthetic data generation and transfer learning.

CRediT authorship contribution statement

Faeze Zakaryapour Sayyad: Conceptualization, Methodology, Software, Validation, Investigation, Data Curation, Writing – original draft, Visualization, Formal analysis. **Irida Shallari:** Writing – review & editing, Supervision. **Seyed Jalaleddin Mousavirad:** Writing – review & editing, Supervision. **Mattias O'Nils:** Writing – review & editing, Supervision. **Faisal Z. Qureshi:** Writing – review & editing.

Declaration of competing interest

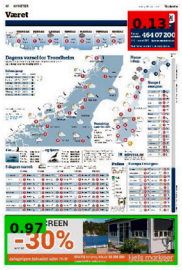
The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Faeze Zakaryapour Sayyad reports financial support was provided by Mid Sweden University. Faeze Zakaryapour Sayyad reports a relationship with Knowledge Foundation (kks.se) within the Industrial graduate school Smart Industry Sweden, Media Research company that includes: employment and funding grants. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

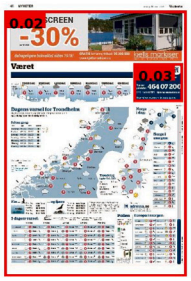
This work was supported in part by The Knowledge Foundation (kks.se) within the Industrial graduate school Smart Industry Sweden and Media Research company.

Data availability

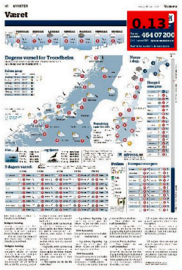
The authors do not have permission to share data.



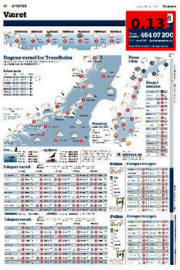
(a) This image shows the original page layout with the advertisement in its initial position at the bottom of the page. This serves as the baseline for our comparisons.



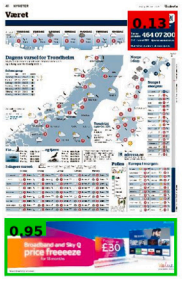
(b) The Ad has been moved from the bottom to the top of the page. This adjustment was made to test if the model's detection accuracy is affected by the Ad's position. The model failed to detect the Ad in this new position, suggesting that it may be overfitted to certain layout patterns.



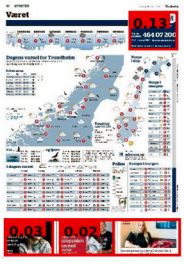
(c) The Ad content was replaced with non-Ad editorial content. The model did not falsely identify this content as an Ad, indicating it can differentiate between typical editorial layouts and advertisements.



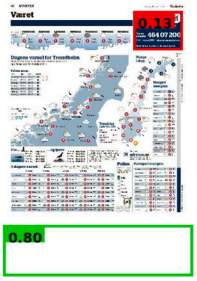
(d) Similarly, when the Ad content was replaced with weather content, the model did not detect it as an Ad, further demonstrating its capability to distinguish between different types of content.



(e) The Ad content was replaced with an Ad from another newspaper. The model successfully detected this new Ad, showing its generalization ability to recognize Ads from different sources.



(f) The Ad content was replaced with editorial content that has a similar layout to an Ad. The model did not detect this as an Ad, indicating its sensitivity to specific Ad features beyond just layout similarity.

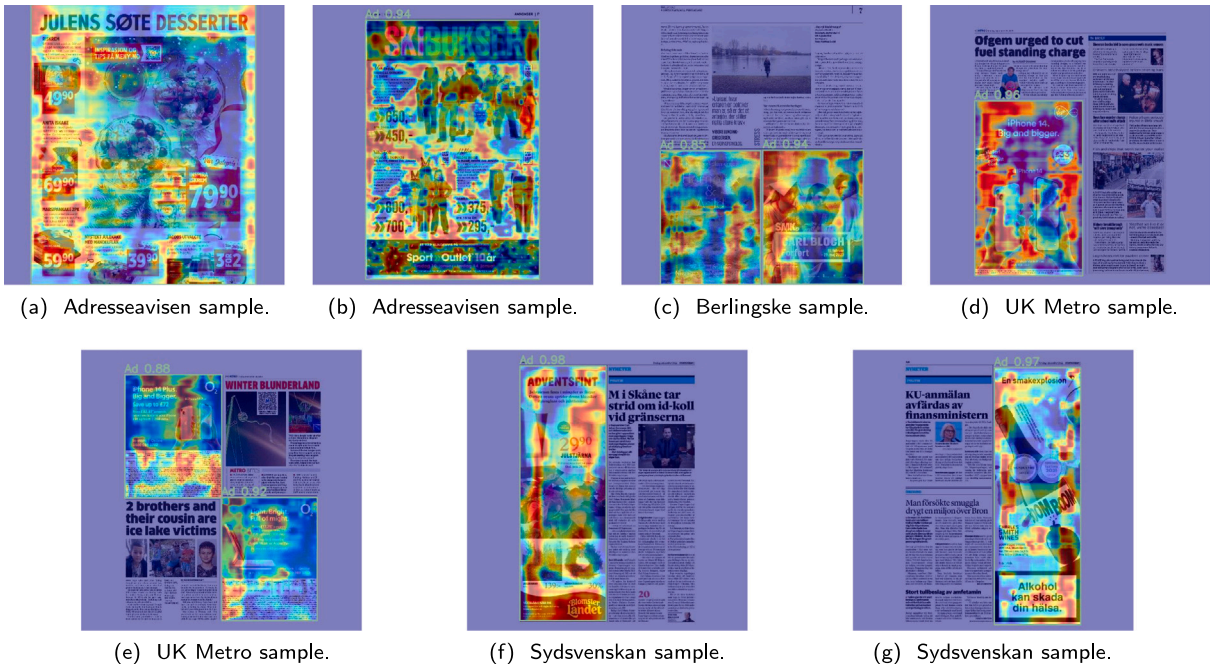


(g) The Ad content was replaced with white space. Surprisingly, the model detected this blank space as an Ad, suggesting that the model might be relying on certain contextual cues that are still present above the white space.

(h) This image and the next one show the Ad moved to different positions on the page. The model's failure to detect the Ad in these new positions highlights the importance of positional information in the current model's detection capability.

(i) Ad location (position) changed.

Fig. 3. Examples of different modifications applied to the images. The bounding boxes indicate detected objects with different confidence levels: high-confidence detections are marked with bold green boxes, while low-confidence detections are marked with bold red boxes. The threshold for distinguishing high and low confidence is set to 0.5. Please not that the results obtained with a threshold of 0.3 were the same as those obtained with a threshold of 0.5.



(a) Adresseavisen sample. (b) Adresseavisen sample. (c) Berlingske sample. (d) UK Metro sample.

(e) UK Metro sample. (f) Sydsvenskan sample. (g) Sydsvenskan sample.

Fig. 4. GradCAM++ heatmaps highlighting YOLOv8 model's attention areas on different newspaper samples.

(a) Original RGB image. (b) Gray Scale. (c) Binary Color. (d) Inverted Color.

Fig. 5. Examples of color variations used to test YOLOv8 model's detection capabilities.

References

Adresseavisen (2022–2023). Adresseavisen. Data extracted over the years 2022 and 2023. (Accessed 26 April 2023).

Aftonbladet (2023). Aftonbladet. Data extracted for the dates 25-31 May 2023. (Accessed 26 June 2023).

Almgren, K., Krishnan, M., Aljanobi, F., & Lee, J. (2018). AD or non-AD: a deep learning approach to detect advertisements from magazines. *Entropy*, *20*(12), 982.

Almutairi, A. (2024). Newspaper elements detection and newspaper pages categorization using CNNs and transformers. *International Journal on Document Analysis and Recognition (IJDAR)*, 1–13.

Almutairi, A., & Almashan, M. (2019). Instance segmentation of newspaper elements using mask R-CNN. In *2019 18th IEEE international conference on machine learning and applications* (pp. 1371–1375). http://dx.doi.org/10.1109/ICMLA.2019.00223.

Álvarez, J., Armenteros, J. C., Torrón, C., Ortega-Martín, M., Ardoiz, A., García, Ó., Arranz, I., Galdeano, Í. n., Garrido, I., Alonso, A., et al. (2024). RADIA–radio advertisement detection with intelligent analytics. arXiv preprint arXiv:2403.03538.

Bansal, A., Mukherjee, P., Joshi, D., Tripathi, D., & Singh, A. P. (2021). Multi-task learning for newspaper image segmentation and baseline detection using attention-based U-net architecture. In *Document analysis and recognition–ICDAR 2021 workshops: lausanne, Switzerland, September 5–10, 2021, proceedings, part II 16* (pp. 440–454). Springer.

Bar, A., Wang, X., Kantorov, V., Reed, C., Herzig, R., Chechik, G., Rohrbach, A., Darrell, T., & Globerson, A. (2022). DETReg: Unsupervised pretraining with region priors for object detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14585–14595).

Barman, R., Ehrmann, M., Clematide, S., Oliveira, S. A., & Kaplan, F. (2021). Combining visual and textual features for semantic segmentation of historical newspapers. *Journal of Data Mining & Digital Humanities*, (HistoInformatics).

Berlingske (2022–2023). Berlingske. Data extracted over the years 2022 and 2023. (Accessed 26 April 2023).

Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision* (pp. 213–229). Springer.

Carvalho, P., Pereira, A., & Viana, P. (2021). Automatic tv logo identification for advertisement detection without prior data. *Applied Sciences*, *11*(16), 7494.

Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision* (pp. 839–847). IEEE.

Chen, Z., Yang, C., Li, Q., Zhao, F., Zha, Z.-J., & Wu, F. (2021). Disentangle your dense object detector. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 4939–4948).

Dhang, S., Zhang, M., & Dev, S. (2024). AdSegNet: a deep network to localize billboard in outdoor scenes. *Signal, Image and Video Processing*, *18*(10), 7221–7235.

Dhang, S., Zhang, M., & Dev, S. (2025). LaBINet-an approach for seamlessly integrating new advertisement into an existing scene. *IEEE Transactions on Artificial Intelligence*.

English, N., Anesetti-Rothermel, A., Zhao, C., Latterner, A., Benson, A. F., Herman, P., Emery, S., Schneider, J., Rose, S. W., Patel, M., et al. (2021). Image processing for public health surveillance of tobacco point-of-sale advertising: Machine learning–based methodology. *Journal of Medical Internet Research*, *23*(8), Article e24408.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580–587).

iTromsø (2023). Itromsø. Data extracted for the dates 25-31 May 2023. (Accessed 26 June 2023).

Jain, P., Taneja, K., & Taneja, H. (2022). Advertisement detection: Image processing and deep learning approach for effective information extraction from online english newspapers. In *Evolutionary computation with intelligent systems* (pp. 39–69). CRC Press.

Jain, P., Taneja, K., & Taneja, H. (2024). Advertisement image classification using deep learning with BERT: A novel approach exploiting textual. *Proceedings of Data Analytics and Management: ICDAM 2023, Volume 2*, *786*, 443.

Jain, P., et al. (2021). Convolutional neural network based advertisement classification models for online english newspapers. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(2), 1687–1698.

Jonsson, P. (2022). A deep learning approach to advertisement detection in newspapers.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980–2988).

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740–755). Springer.

Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., & Zhang, L. (2022). DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *International conference on learning representations*.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012–10022).

Liu, D., Ni, J., Lin, X., & Shen, X. (2020). Transparent and accountable vehicular local advertising with practical blockchain designs. *IEEE Transactions on Vehicular Technology*, *69*(12), 15694–15705.

Lu, X., Li, B., Yue, Y., Li, Q., & Yan, J. (2019). Grid r-cnn. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Lyu, C., Zhang, W., Huang, H., Zhou, Y., Wang, Y., Liu, Y., Zhang, S., & Chen, K. (2022). RTMDet: An empirical study of designing real-time object detectors. arXiv: 2212.07784.

Madi, B., Alaasam, R., Droby, A., & El-Sana, J. (2021). Advertisement extraction using deep learning. In *Document analysis and recognition–ICDAR 2021 workshops: lausanne, Switzerland, September 5–10, 2021, proceedings, part II 16* (pp. 81–97). Springer.

Metro (2022–2023). Metro. Data extracted over the years 2022 and 2023. (Accessed 26 April 2023).

Morera, Á., Sánchez, Á., Moreno, A. B., Sappa, Á. D., & Vélez, J. F. (2020). SSD vs. YOLO for detection of outdoor urban advertising panels under multiple variabilities. *Sensors*, *20*(16), 4587.

Nazeri, M., Wang, J., Payandeh, A., & Xiao, X. (2024). Vanp: Learning where to see for navigation with self-supervised vision-action pre-training. In *2024 IEEE/RSJ international conference on intelligent robots and systems* (pp. 2741–2746). IEEE.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779–788).

Reis, D., Kupec, J., Hong, J., & Daoudi, A. (2023). Real-time flying object detection with YOLOv8. arXiv preprint arXiv:2305.09972.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, *28*.

Rondan, N., Juayek, M., Joskowicz, J., Sotelo, R., Quincke, S., Patrone, A., Aguerre, M., & Gonzalez, G. (2025). Automatic detection of TV commercial breaks. In *2025 IEEE international conference on consumer electronics* (pp. 1–6). IEEE.

Siddiqui, M. A., Hyder, M. F., Mukarram, M., et al. (2021). TV ad detection using the base64 encoding technique.. *Engineering, Technology & Applied Science Research*, *11*(5).

Sydsvenskan (2022–2023). Sydsvenskan. Data extracted over the years 2022 and 2023. (Accessed 26 April 2023).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Wang, C.-Y., Liao, H.-Y. M., Wu, Y.-H., Chen, P.-Y., Hsieh, J.-W., & Yeh, I.-H. (2020). Cspnet: A new backbone that can enhance learning capability of CNN. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 390–391).

Wardana, M. Z. P., & Wibowo, M. E. (2023). Audio-visual cnn using transfer learning for tv commercial break detection. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *17*(3), 291–300.

Wightman, R., Touvron, H., & Jégou, H. (2021). Resnet strikes back: An improved training procedure in timm. arXiv preprint arXiv:2110.00476.

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1492–1500).

Xie, Y., Nie, Y., Lundgren, J., Yang, M., Zhang, Y., & Chen, Z. (2024). Cervical spondylosis diagnosis based on convolutional neural network with X-ray images. *Sensors*, *24*(11), 3428.

Yang, S., Liu, Y., Liu, Z., Xu, C., & Du, X. (2024). Enhanced vehicle logo detection method based on self-attention mechanism for electric vehicle application. *World Electric Vehicle Journal*, *15*(10), 467.

Yang, Z., Pei, W., Chen, M., & Yue, C. (2022). Wtagraph: Web tracking and advertising detection using graph neural networks. In *2022 IEEE symposium on security and privacy* (pp. 1540–1557). IEEE.

Ye, K., & Kovashka, A. (2018). Advise: Symbolism and external knowledge for decoding advertisements. In *Proceedings of the European conference on computer vision* (pp. 837–855).

Zakaryapour Sayyad, F., Shallari, I., Mousavirad, S. J., & O'Nils, M. (2024). Model evaluation and selection for robust and efficient advertisement detection in print media. In *International conference on advances in computing and data sciences* (pp. 211–224). Springer.

Zhang, H., Huang, H., Zhao, P., Zhu, X., & Yu, Z. (2024). CENN: Capsule-enhanced neural network with innovative metrics for robust speech emotion recognition. *Knowledge-Based Systems*, *304*, Article 112499.

Zhao, P., Liu, X., Yue, Z., Zhao, Q., Liu, X., Deng, Y., & Wu, J. (2024). DiGAN breakthrough: Advancing diabetic data analysis with innovative GAN-based imbalance correction techniques. *Computer Methods and Programs in Biomedicine Update*, *5*, Article 100152.

Zhao, S., Zhou, D., Wang, H., Chen, D., & Yu, L. (2025). Enhancing student academic success prediction through ensemble learning and image-based behavioral data transformation. *Applied Sciences*, *15*(3), 1231.

Zou, Z., Chen, K., Shi, Z., Guo, Y., & Ye, J. (2023). Object detection in 20 years: A survey. *Proceedings of the IEEE*.