separations

MDPI

Article

Comprehensive Quantitative Analysis of Coal-Based Liquids by Mask R-CNN-Assisted Two-Dimensional Gas Chromatography

Huan-Huan Fan [1,2], Xiang-Ling Wang [1], Jie Feng [1,2,3,*] and Wen-Ying Li [1,2,3]

[1] State Key Laboratory of Clean and Efficient Coal Utilization, Taiyuan University of Technology, Taiyuan 030024, China; fanhuanhuan0182@link.tyut.edu.cn (H.-H.F.); m15886531716@163.com (X.-L.W.); ying@tyut.edu.cn (W.-Y.L.)
[2] Shanxi Research Institute of Huairou Laboratory, Taiyuan 030032, China
[3] Beijing Huairou Laboratory, Beijing 101499, China
[*] Correspondence: fengjie@tyut.edu.cn; Tel.: +86-0351-6018453

Abstract: A comprehensive understanding of the compositions and physicochemical properties of coal-based liquids is conducive to the rapid development of multipurpose, high-performance, and high-value functional chemicals. However, because of their complex compositions, coal-based liquids generate two-dimensional gas chromatography (GC × GC) chromatograms that are very complex and very time consuming to analyze. Therefore, the development of a method for accurately and rapidly analyzing chromatograms is crucial for understanding the chemical compositions and structures of coal-based liquids, such as direct coal liquefaction (DCL) oils and coal tar. In this study, DCL oils were distilled and qualitatively analyzed using GC × GC chromatograms. A deep-learning (DL) model was used to identify spectral features in GC × GC chromatograms and predominantly categorize the corresponding DCL oils as aliphatic alkanes, cycloalkanes, mono-, bi-, tri-, and tetracyclic aromatics. Regional labels associated with areas in the GC × GC chromatograms were fed into the mask-region-based convolutional neural network's (Mask R-CNN's) algorithm. The Mask R-CNN accurately and rapidly segmented the GC × GC chromatograms into regions representing different compounds, thereby automatically qualitatively classifying the compounds according to their spots in the chromatograms. Results show that the Mask R-CNN model's accuracy, precision, recall, F1 value, and Intersection over Union (IoU) value were 93.71%, 96.99%, 96.27%, 0.95, and 0.93, respectively. DL is effective for visually comparing GC × GC chromatograms to analyze the compositions of chemical mixtures, accelerating GC × GC chromatogram interpretation and compound characterization and facilitating comparisons of the chemical compositions of multiple coal-based liquids produced in the coal and petroleum industry. Applying DL to analyze chromatograms improves analysis efficiency and provides a new method for analyzing GC × GC chromatograms, which is important for fast and accurate analysis.

Keywords: direct coal liquefaction oils; chemical composition; deep learning; pattern recognition

check for updates

Academic Editor: Victoria Samanidou

Received: 2 January 2025
Revised: 20 January 2025
Accepted: 20 January 2025
Published: 24 January 2025

Citation: Fan, H.-H.; Wang, X.-L.; Feng, J.; Li, W.-Y. Comprehensive Quantitative Analysis of Coal-Based Liquids by Mask R-CNN-Assisted Two-Dimensional Gas Chromatography. *Separations* **2025**, *12*, 22. https://doi.org/10.3390/separations12020022

1. Introduction

Coal-based liquids encompass a range of high-quality coal-derived raw materials, including direct coal liquefaction (DCL) oils and coal tar, and processed products, such as naphtha, white oil, diesel fuel, and aviation kerosene [1], and are produced by breaking the chemical structure of coal macromolecules while retaining certain characteristic chemical structures, such as coal's inherent ring structures, generating complex mixtures suitable for further processing to high-performance products. This complexity presents challenges for analyzing coal-based liquids. To improve product quality, optimize production, reduce production costs, and ensure that products meet market demands, industrially produced oils must be rapidly analyzed and verified. A comprehensive understanding of the compositions and physicochemical properties of coal-based liquids is essential for efficiently utilizing their properties and swiftly developing multipurpose, high-performance, and high-value products. Consequently, a convenient method must be developed for rapidly analyzing components of coal-based liquids.

Coal-based liquids contain hundreds or even thousands of components. With the development of two-dimensional (2D) gas chromatography (GC × GC) [2], researchers have increasingly used it to analyze complex samples [3] because of its high resolution, peak capacity, and sensitivity compared with those of traditional unidimensional (1D) GC [4–10]. However, the GC × GC chromatograms of complex samples render qualitative analysis and processing cumbersome and time-consuming, and analysts must rely on their experience to interpret mass spectra and may overlook certain details, leading to subjective results [11]. Pattern recognition and classification tools are essential for compliance with industrial standards [12]. An appropriate method can effectively improve efficiency in analyzing GC × GC chromatograms, where spots corresponding to chemical compounds are arranged according to specific patterns, enabling thousands of compounds to be classified using Visual Basic Scripts [13–15], which can substantially shorten analysis times and improve classification accuracy [16] for complex samples. One method involves constructing spatial polygonal maps in GC × GC chromatograms, which classify structurally similar compounds into the same group. Kehimkar [17] used comprehensive two-dimensional gas chromatography–mass spectrometry (GC × GC–MS) to analyze jet fuels and categorized the compounds in the GC × GC chromatogram into five groups. Liu [18] used the Computer Language for Identifying Compounds, which applies an interactive template function for matching mass fragmentation features and retention times, to construct spatial polygons and classify coal tar into 10 categories; however, typically represented templates must be pre-categorized based on experience. Furthermore, templates obtained using these analysis methods are fixed and inflexible, leading to potential template failure when testing conditions vary, and these methods lack versatility and are time consuming. Therefore, a universal, effective, rapid, and accurate method must be developed for efficiently processing and analyzing GC × GC chromatograms.

Deep learning (DL), a branch of machine learning (ML), is known for its ability to automatically extract complex features and replicate human visual capabilities in image processing [19]. DL is essential for image processing tasks, including restoration, enhancement, segmentation, and feature extraction [20–24]. Because numerous GC × GC chromatograms contain regular information, such as the "tile effect", DL can be utilized to fully extract this information, enabling automatic learning and pattern recognition through advanced algorithms. A mask-region-based convolution neural network (Mask R-CNN) is a DL model specifically designed for detection and instance segmentation tasks, and it can detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance [25,26]. Furthermore, in computer vision, Mask R-CNN is an important DL application utilizing deep CNNs to precisely localize and segment objects in images. This approach can reduce manual intervention and enhance both efficiency and accuracy in chromatogram analysis. ML techniques have been applied to analyze the distributions and compositions of the products of co-pyrolyzed biomass and coal [27]. Additionally, researchers have developed comparative visualization methods, such as using differential images, to analyze chemical compositional differences [28]. However, these methods require consistent sample testing conditions, are limited to comparing substances' chromatograms, and are not conducive for rapidly analyzing different chromatograms. Because of the wide variety of oils and products derived from coal-based liquids, Mask R-CNN has been used to analyze chromatograms and improve the efficiency of compound spot identification. Because their chemical compositions vary substantially, coal-based liquids must be accurately and rapidly analyzed for further processing. However, in the coal chemical industry, the current development of Mask R-CNN for application to image pattern recognition is insufficient. Moreover, to the best of our knowledge, few reports on the combination of GC × GC chromatogram analysis with Mask R-CNN are available in the literature.

Therefore, in this study, the composition of DCL oils was analyzed using Mask R-CNN for recognizing patterns in GC × GC chromatogram spots corresponding to different compounds. By combining established patterns and analysts' experience, unknown spots can be identified, and the analysis accuracy can be enhanced. Mask R-CNN provides a fast and convenient method for analyzing coal-based liquids. The performance of the Mask R-CNN model in the task of identifying the GC-GC chromatograms of DCL oils and classifying their different compositions is excellent. The Mask R-CNN model's accuracy, precision, recall, F1 value, and IoU value were 93.71%, 96.99%, 96.27%, 0.95, and 0.93, respectively. By using Mask R-CNN to recognize the GC × GC chromatograms of DCL, the components in the chromatograms are classified into six categories, which can compare the compositional differences and content differences of different oils. Overall, DL offers a new perspective for analyzing GC × GC chromatograms, enabling more efficient and accurate data processing in chemical analysis through automated feature learning and pattern recognition. DL solves the problem of long analysis time in GC × GC chromatograms, improves analyze efficiency, and makes spectral chromatograms analysis more comprehensive and systematic. DL technological convergence promotes multidisciplinary cross-collaboration among chemistry, computer science, artificial intelligence, and other disciplines to drive innovation in analytical chemistry research.

2. Materials and Methods

2.1. Reagents and Materials

A distillation column unit (Tianjin Aozhan Chemical Technology Company, Tianjin, China) was used to separate the DCL oils. This study involved 36 DCL oil samples, comprising 6 light-fraction oils, 2 naphthas, 1 white oil, and 27 distillates, derived from 2 distinct oils, which can be categorized into 16 and 11 groups, respectively. Details are shown in Table 1.

Table 1. 36 sample oils investigated in this study.

Sample	Number	Main Composition
light-fraction oils	6	cycloalkanes, polycyclic aromatics
naphtha	2	polycyclic aromatics
white oil	1	cycloalkanes
distillates	27	complex chemical composition

For subsequent use, an oil solution (0.01 mol/L) was prepared using analytically pure methylene chloride (Shanghai Aladdin Biochemical Technology Co., Ltd., Shanghai China) as the solvent, which was employed as received without further treatment or purification.

2.2. GC × GC–MS Analytical Methods

GC × GC coupled with MS and a flame ionization detector (FID) is an advanced analytical technique that allows for simultaneous qualitative analysis by MS and quantitative analysis by FID. A Shimadzu QP2020 instrument (Shimadzu, Kyoto, Japan) was used for this purpose. The GC × GC–MS/FID conditions were as follows: Unidimensional and two-dimensional separations were performed using a DB-1 column (15 m × 0.25 mm × 0.25 µm) and a BPX-50 column (2.75 m × 0.1 mm × 0.1 µm), respectively. The initial column temperature was set at 60 °C and programmed to increase at 3 °C/min to 280 °C and held there for 5 min. The injection port temperature was maintained at 280 °C, and the carrier gas was high-purity helium (99.99% vol). The modulation period was set at 6 s.

The MS conditions included a solvent delay of 6 min, an electron ionization (EI) source at 250 °C, and electron bombardment at 70 eV. The mass spectrometer scanned a range of m/z values from 45 to 400 amu at 50 Hz. The shunt ratio was set at 30:1. The injection volume was 0.6 mL. Data acquisition and analysis were performed using Mass Insight and GC Image 2.7 software (GC Image Limited Liability Company (LLC) Lincoln, NE, USA).

2.3. Mask R-CNN

The training hardware environment comprised a 15.5 GB memory, an Intel i7-9700K processor, an Nvidia Quadro P620 graphics card, and an Ubuntu 18.04.6 long-term support (LTS) operating system, and Python was utilized as the programming language. An environment was created using the Conda package manager, within which the LabelMe and Mask R-CNN packages were installed. LabelMe was used to annotate each region in the GC × GC chromatograms of the DCL oils with polygons, supplying the labels required for model training.

2.3.1. Architecture of the Mask R-CNN Model

In the GC × GC chromatograms recognition, the Mask R-CNN's architecture as shown in Figure 1. The initial image processing and analysis step is the input of images, which are preprocessed to ensure they are suitable for network model processing. This includes operations, such as resizing images and normalizing pixel values, to maintain consistency in the input data. Next, the model extracts feature at different levels from the images. These features are then merged to form a feature map that contains rich information about various targets in the images. The feature map is crucial for the model's detection and segmentation tasks. Finally, the model uses a segmentation layer to identify the targets in the image and precisely segments each target. See Appendix A for more details on the construction of the Mask R-CNN's architecture.

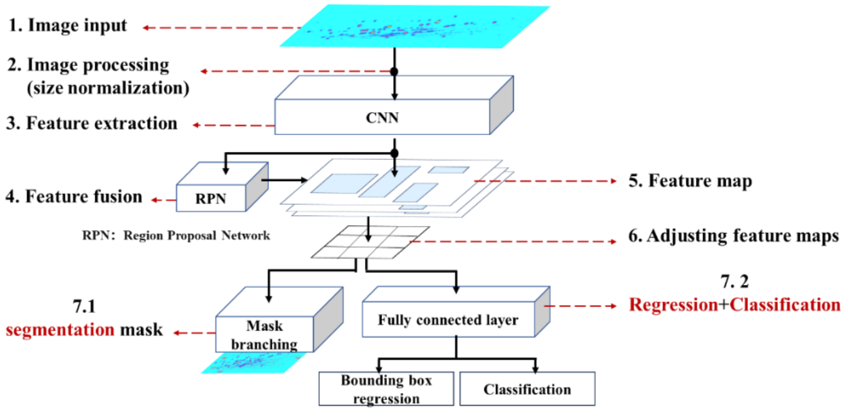

Figure 1. Mask R-CNN's architecture for GC × GC chromatograms recognition.

In the recognition of GC × GC chromatograms, the architecture of the Mask R-CNN is depicted in Figure 1. The process commences with the input of chromatogram images, which undergo initial preprocessing to align with the network model's requirements. This preprocessing encompasses resizing the images and normalizing the pixel values, ensuring uniformity in the input dataset. Subsequently, the model delves into multi-level feature extraction from these images. The extracted features are amalgamated to construct a comprehensive feature map, replete with detailed information on the diverse targets within the images. This feature map is pivotal for the model's proficiency in detection and segmentation tasks. Concluding the process, a segmentation layer is employed to pinpoint and accurately demarcate each target within the image. For an in-depth exploration of the Mask R-CNN's architectural framework, refer to Appendix A.

2.3.2. Algorithm of the Mask R-CNN Model

The Mask R-CNN's algorithm is utilized for recognizing patterns, preprocessing chromatograms by normalizing their sizes, and generating segmentation masks through bounding box regression and region of interest (ROI) evaluation. The Mask R-CNN's algorithm was primarily trained using the 36 previously analyzed chromatograms of the DCL oil, which were categorized into six main regions labeled from A to F.

For model testing, 1 chromatogram was randomly selected, while the remaining 35 were used for model training. The Mask R-CNN algorithm's parameters, which were optimized for segmenting the GC × GC chromatogram's spots (corresponding to the DCL oils' components) into regions, included a ResNet101 backbone network and an image size of 384 × 384; the learning rate was set at 0.001, and a total of 150 training rounds were conducted. Each iteration cycle involved 100 training and 30 validation steps, with bounding box sizes configured at 8 × 6, 16 × 6, 32 × 6, 64 × 6, and 128 × 6. The training hardware environment included 15.5 GB of memory, an Intel i7-9700K processor, an Nvidia Quadro P620 graphics card, and an Ubuntu 18.04.6 LTS operating system, and Python was used as the programming language.

2.3.3. Training of the Mask R-CNN Model

The ResNet neural network was first pretrained using training set samples to extract features from GC × GC chromatograms. Then, mask and classifier branches were added to train the network model's parameters using the optimized training set samples. After multiple training and transfer-learning adjustment iterations, the model was optimized. Finally, validation set samples were used to verify and further adjust the model's accuracy. Model training focused on loss function convergence and training set recognition as well as key parameters.

2.3.4. Loss Function of the Mask R-CNN Model

A non-negative real-valued loss function, which measures the difference between the model-predicted and actual values, plays a crucial role in the Mask R-CNN model. During model training, the loss function is typically minimized to improve the model's prediction accuracy. An appropriate loss function is essential for improving the model's prediction accuracy because the loss function determines how the model learns from the data.

3. Results and Discussion

3.1. Qualitative Analysis of the DCL Oils

The analysis of the DCL oils revealed notable compositional shifts with rising boiling points. With increasing boiling point, the carbon count of the compounds in the fractions gradually increased. The compound categories evolved from initially comprising shorterchain and monocyclic alkanes, benzene, and some monocyclic aromatic hydrocarbons, to comprising longer-chain and tetracyclic alkanes, tricyclic aromatics, tetracyclic aromatics, and other similar compounds. The structurally similar compounds were arranged in orderly rows, resulting in the formation of a tiling effect.

Figure 2 shows the GC × GC chromatogram of the DCL oil and clearly reveals the complexity of the DCL oil′ composition, and GC × GC can separate the substances in the DCL oil. The spots of the compounds are arranged in an orderly manner in the chromatogram. In GC × GC, the nonpolar column is connected to a second polar column; this setup can generate structured chromatograms, where spots corresponding to the different compounds are arranged according to their chemical groups and the number of carbon atoms they contain [29] and where the substances with similar chemical properties are distributed in bands. In GC × GC chromatograms, the spots corresponding to all the compounds in a sample are arranged in a certain pattern; therefore, the compounds can be classified as aliphatic alkanes, cycloalkanes, mono- and polycyclic aromatic compounds according to the spot arrangement from bottom to top. Thus, in the contour map, the spots corresponding to the different substances are categorized into distinct regions. Because the content of N- and S-containing compounds in the DCL oil is low at 0.82%, these compounds were excluded from the regional segmentation.

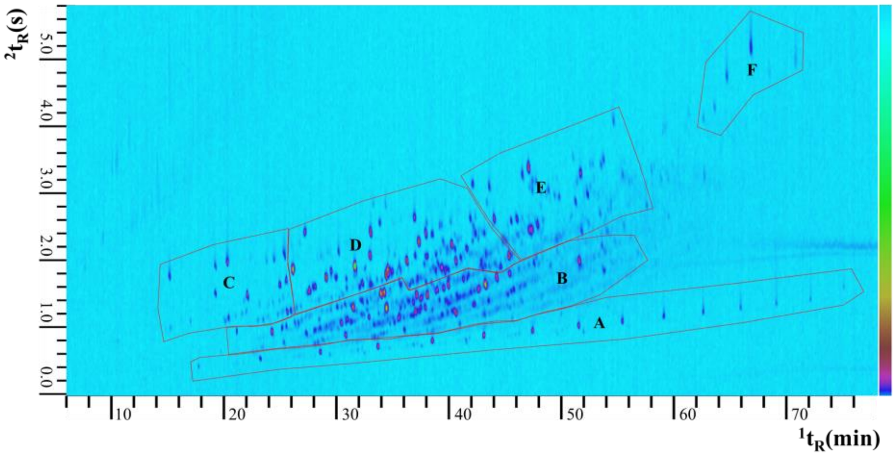

Figure 2. The six elution regions in the GC × GC chromatogram of the DCL oil′ component (A: aliphatic alkanes, B: cycloalkanes, C: monocyclic aromatics, D: bicyclic aromatics, E: tricyclic aromatics, F: tetracyclic aromatics).

To ensure that the recognition model's accuracy was not compromised by the data quantity, the categorization of the DCL oils' substances was designed to avoid using too much data. These compounds can be classified into six categories comprising aliphatic alkanes, cycloalkanes (including monocyclic, bicyclic, tricyclic, and tetracyclic), monocyclic (including oxygen-containing substances, such as phenolics, esters, and carboxylic acids), bicyclic, tricyclic, and tetracyclic aromatics. These six elution regions, labeled from A to F and shown in Figure 2, were established to provide data for pattern recognition across the various zones in the DCL oils' GC × GC chromatograms.

In Figure 2, spots corresponding to aliphatic alkanes are at the base of the chromatogram. Spots corresponding to cycloalkanes, which are more polar than their acyclic counterparts bearing the equivalent number of carbon atoms, are above the spots corresponding to the aliphatic alkanes. Spots corresponding to aromatic compounds appear at the top of the chromatogram because aromatic compounds are more polar than both aliphatic alkanes and cycloalkanes. The aromatics substances are ordered from left to right, according to increasing boiling points and longer elution times from the 1D column, as follows: benzene, naphthalene, anthracene, phenanthrene, and pyrene. Clearly, this sample contains many hundreds of components, rendering the manual analysis of each chromatographic spot extremely laborious for analysts.

3.2. Pattern Recognition GC × GC Chromatograms of the DCL Oils

GC × GC chromatograms represent compounds as spots in three-dimensional(3D) space, with retention times on the *x*- and *y*-axis and the signal intensity on the *z*-axis. Figure 3 shows the 3D GC × GC chromatogram of components of DCL oil: the *x*-axis represents the retention time of the first dimensional column, the *y*-axis represents the retention time of the second dimensional column, and the *z*-axis represents the signal intensity. Each compound corresponds to a point in the chromatogram, the position of which is determined by the retention time on both columns, while the signal intensity is indicated by the color of the point, with darker colors indicating greater signal intensity. In the 2D chromatogram, these compounds with the same retention time in the 1D dimensional column are further separated in the second dimensional column, so that their points in the 2D chromatogram will have different positions in the *y*-axis direction, and the color of the points with different signal intensities is also different, so that these compounds, which are originally difficult to be distinguished in 1D chromatogram, can be distinguished, and at the same time can be distinguished from the background noise. The signal intensities of the compounds are usually different because the signal strength of the background noise is usually weak, and its corresponding points are lighter in color, which is significantly different from the point colors of the compounds. The blue is the background and the color of each spot indicates its signal intensity, forming an image comprising pixel dots of various colors representing different signal intensities. This visual representation aids in distinguishing between compounds and background noise [11,30,31].

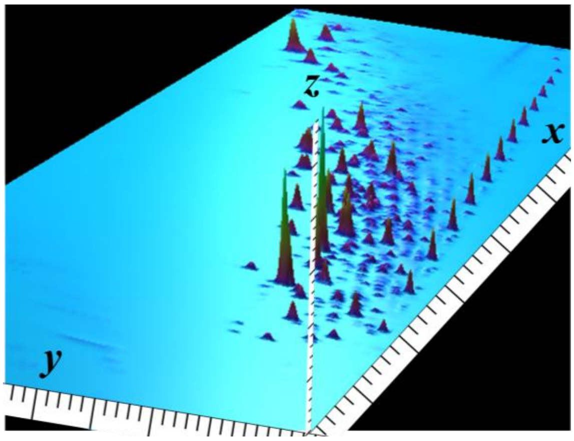

Figure 3. The 3D GC × GC chromatogram of components of DCL oil.

GC × GC orthogonally separates oil components by distributing substances possessing similar structural properties into distinct bands. The DL model learns and extracts key information in a manner like that of its capacity to discriminate between the minute differences in various human facial features, and this information is combined for effective recognition and classification.

In the GC × GC chromatograms, the retention time is directly related to a compound's boiling point and polarity. Chromatogram data are organized in a pixelated format, with 1D and GC × GC separation times along the horizontal and vertical axes, respectively. Compounds are further arranged according to their carbon atom counts and chemical groups, generating a structured chromatogram. Spots corresponding to similar substances are regularly distributed, and their spectral positions are interrelated, which facilitates the classification of complex samples into distinct families. Furthermore, the positional information of the different spots enables the identification of distinct regions using DL. This approach facilitates the recognition of areas where spots corresponding to various compounds are located. The pattern recognition model utilizes the algorithm of Mask R-CNN to automatically divide the DCL oils' GC × GC chromatograms into different regions corresponding to their different components.

3.2.1. Positions of Compound-Related Spots in Chromatograms

GC × GC chromatograms can be treated as multichannel digital images, where each data point corresponds to a single pixel. The DCL oils' chromatogram utilized in DL image recognition does not inherently contain the structural information of compounds corresponding to individual spots. The correlation between the MS data of the individual peaks and the data in the INST MS library facilitates MS-based compound identification. The positions of the different compounds' spots are interrelated, enhancing the identification. Parameters of the GC × GC method can shift the spots of the same compounds in the chromatograms of different DCL oils.

Even for the same compound, the position, shape, and other spot characteristics can vary across different GC × GC chromatograms. As shown in Figure 4, spots corresponding to substances a, b, c, and d are identified as belonging to 1-propenyl-4-methylbenzene; 1-ethylene-2,4-dimethylbenzene; 1,2,3,4-tetrahydronaphthalene (THN); and naphthalene, respectively. Although these compounds' spots are represented in different styles in both chromatograms, they all appear as pixels that are more intense than the background because of the rasterization of the peak signals into pixelated dots. In a chromatogram, each peak represents the elution signal of a compound. To more accurately calculate the peak area, the peak can be divided into multiple subsections or slices. The signal intensity of each slice is measured individually, and then the signal intensities of these slices are summed to obtain the total area of the entire peak. This method helps to improve the accuracy of peak area measurements, especially in cases of peak overlap or baseline drift [28,32,33]. The analysis of GC × GC chromatograms is primarily focused on the elimination of background noise and the normalization of retention time shifts. These preprocessing steps can diminish the influences of extraneous variables (including background noise and the normalization of retention time shifts), thereby facilitating pixel-based analysis that does not require data integration or peak deconvolution. The initial compound identification and peak annotation are conventionally accomplished by referencing NIST MS library, and numerous algorithms can query extensive MS libraries, e.g., NIST MS database, MassBank, Wiley MS database, Sadlter MS database. However, although such algorithms may also generate error messages during searches, this likelihood can be substantially diminished by implementing DL-enhanced library search methods [34].

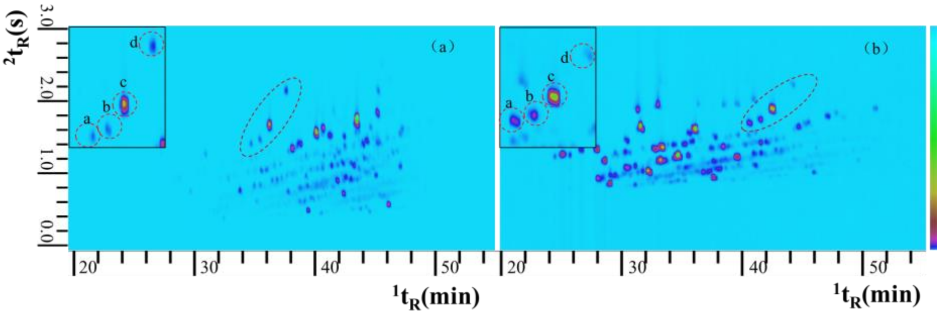

Figure 4. Local areas in GC × GC chromatogram. (**a**) Coal tar chromatogram; (**b**) DCL oils' chromatogram (a: 1-propenyl-4-methylbenzene b: 1-ethylene-2,4-dimethylbenzene, c: 1,2,3,4-tetrahydronaphthalene (THN), and d: naphthalene).

Variations in spot morphology do not necessarily indicate differences in compound structure. The GC × GC dataset often shows inconsistencies in spot retention times and shapes, which may stem from uncontrolled chromatographic variations unrelated to the chemical compositions of the samples. These data inconsistencies and inherent complexities present key challenges to the computer-based visualization and analysis of GC × GC chromatograms. As shown in Figure 4, spots corresponding to THN are adjacent to those corresponding to naphthalene, indicating a correlation between their spot positions in the chromatogram. The structured chromatograms contain rich information, enabling the analysis of spot positions to deduce the distributions of specific compounds across various oils.

3.2.2. Mask R-CNN Segmentation Results

DL can be used to recognize spots in GC × GC chromatograms to rapidly analyze chemical compositions. In image recognition, segmentation is an important task [35]. Fully convolutional network enables the Mask R-CNN to identify the location and category of objects in images when processing multiple targets and precisely segment the contours of each object. For all the oil samples, the substance compositions and distributions were clarified through the detailed analysis of the chromatograms. In order to compare the differences between Mask R-CNN image recognition and manual analysis of GC × GC spectrograms of DCL oil, and to assess the effectiveness of GC × GC chromatogram segmentations in Mask R-CNN, Figure 2 was selected for image segmentation. The result is shown in Figure 5.

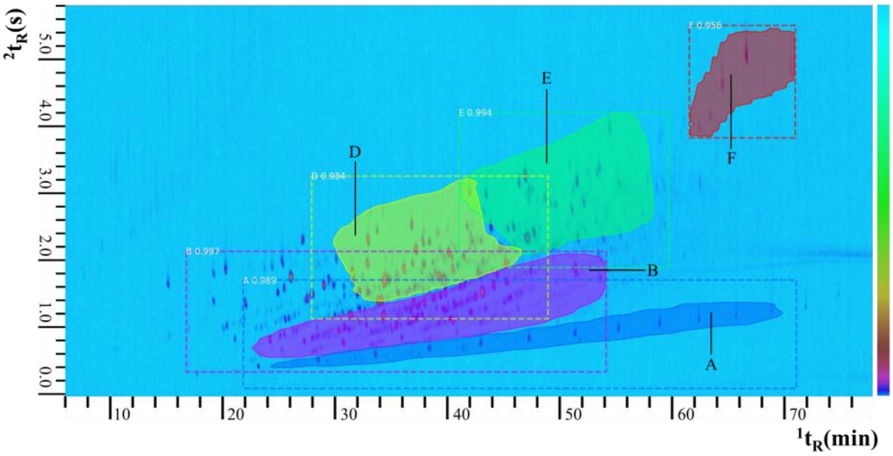

Figure 5. Mask R-CNN segmented regions of GC × GC chromatogram (A: aliphatic alkanes, B: cycloalkanes, D: bicyclic aromatics, E: tricyclic aromatics, F: tetracyclic aromatics).

The results show that the Mask R-CNN accurately identified and outlined regions A, E, and F. Regions A and F, which contain more discernible chromatographic spots, correspond to simpler substance compositions, whereas regions B and D, although corresponding to more complex substance compositions than regions A and F, still exhibit relatively distinct peaks, allowing for improved segmentation, especially considering the smoother edges of region E, as observed during model training. However, in regions B and D, pattern recognition was slightly less effective, mainly because of the high density of peak spots and rugged edges of region labels during model training, which compromised the segmentation quality in these regions compared with that in regions A, E, and F. During image segmentation, regions may be incompletely segmented due to image quality, noise, or algorithm limitations, leading to blocked or unclear areas. Overlapping regions can occur, where a pixel or area belongs to multiple objects simultaneously, affecting segmentation accuracy and subsequent analysis. Additionally, segmented regions may be incomplete with incorrectly recognized and cut boundaries, causing segmentation results to mismatch actual objects. Although the model recognizes patterns in all the regions except for monocyclic aromatics (region C), it still encounters issues, such as local blocking, overlapping, and incomplete cutting, at the edges of some regions. The failure to identify and segment region C is primarily attributed to the low content of phenolic compounds in the DCL oils and the limited number of training chromatograms featuring region C. In region F, the peak spots are more regular, enabling the model to learn this region's features more effectively than those of region C. To enhance the Mask R-CNN's identification accuracy and prediction precision, the dataset should be expanded to include additional chromatograms. The Mask R-CNN's learning rate, number of training epochs, and architecture were fine-tuned to optimize the model's recognition accuracy for features in the GC × GC chromatograms of diverse DCL oils, enabling the model to accurately discern unique features.

3.2.3. Expanded Chromatogram Segmentation Results

The performance of the DL model depends on the quality and quantity of the training data. The existing dataset comprising 35 GC × GC chromatograms falls short of the training requirements, necessitating the expansion of the dataset by increasing the number of available chromatograms. Because of the limited availability of oil samples, image processing techniques, such as rotation, cropping, and shifting, were employed to augment the dataset. Utilizing GC Image 2.7 software, the chromatograms were adjusted and resampled multiple times. The training dataset was expanded to 143 chromatograms through color, saturation, spot size, and position modifications. The steps of chromatography processing, including import, pre-processing, and generation, were documented in detail. This ensured transparent and reproducible chromatography processing. The same chromatography set was processed several times to ensure that the results of each generated chromatogram were consistent and to achieve reproducible analysis of the chromatography. The generated chromatograms were compared with known standard chromatograms to ensure the accuracy of the results. Statistical analysis was performed on the generated chromatograms, including calculation of the mean and standard deviation, to assess the stability and reliability of the chromatography.

This enhancement strategy was implemented by generating new images from multiple sampling iterations, thereby improving the model's accuracy. The same parameters were used to retrain the Mask R-CNN model with the expanded dataset. Then, the model was tested using the expanded dataset, including the DCL oils' chromatograms from the original dataset, and the results are shown in Figure 6.

The Mask R-CNN model of DL rapidly analyzed multiple chromatograms, enabling spectral analysis within 10 s and eliminating the need to manually analyze the chromatograms of individual compounds. It shows enhanced performance in chromatogram segmentation and compound identification, notably by identifying region C compounds that were not identified in Figure 5. This enhances the ability of Mask R-CNN to recognize compounds in chromatograms by utilizing the extended database. This enhancement is attributed to the expanded database, which allows the model to reach a more diverse set of compound features, thus improving the accuracy and comprehensiveness of recognition. This expansion of the database significantly enhances the model's ability to discriminate between different compounds in the chromatogram.

Furthermore, the results in Figure 6 are highly comparable to the data obtained from manual analysis (Figure 2). This close correlation indicates that the Mask R-CNN model, with an expanded database, has high accuracy in analyzing chromatograms. The validity and reliability of the model is further confirmed by the small difference between the segmentation results in Figures 2 and 6. The Mask R-CNN model performs excellently in analyzing different regions in the chromatogram after the database has been expanded.

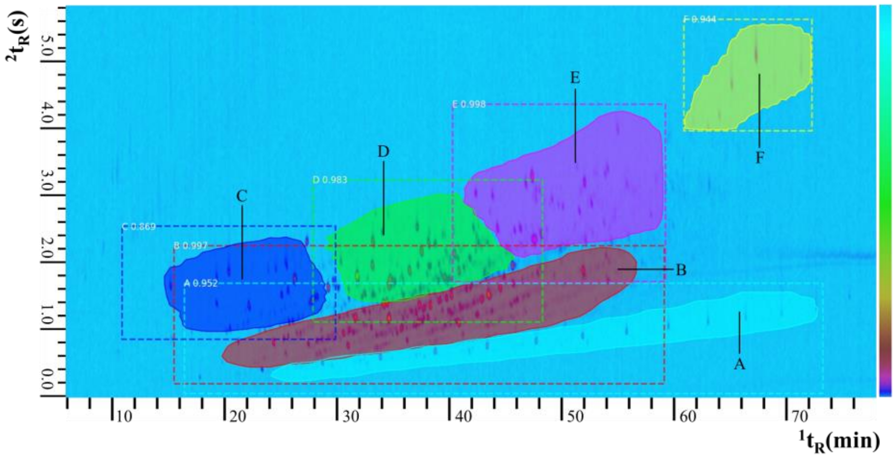

Figure 6. GC × GC chromatograms segmented using the Mask R-CNN model trained using more expanded chromatograms (A: aliphatic alkanes, B: cycloalkanes, C: monocyclic aromatics, D: bicyclic aromatics, E: tricyclic aromatics, F: tetracyclic aromatics).

By using Mask R-CNN to recognize the GC × GC chromatograms of DCL oils, the components in the chromatograms are classified into six categories. This method can compare compositional differences between oils and roughly compare their content differences. Mask R-CNN is a deep learning-based instance segmentation algorithm that accurately identifies and segments different chromatographic peaks in complex chromatograms. In this way, researchers can more accurately analyze the individual components in the chromatograms to gain insight into the chemical composition and content variations of different oils. Mask R-CNN can assist researchers in the field of chemical analysis to quickly and accurately resolve complex chromatograms. The application of this technique provides a new and more efficient method for analyzing coal-based liquid oils.

3.2.4. Evaluation Indicators of the Mask R-CNN

Loss function is a metric used to measure the difference between the model-predicted and actual values and can quantitatively measure the model's accuracy. The loss function curve for the Mask R-CNN model trained using the expanded dataset is shown in Figure 7.

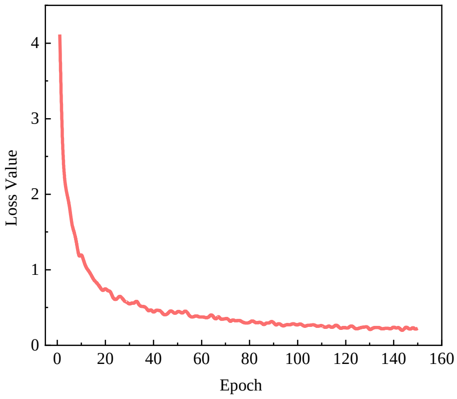

Figure 7. Loss function curve for the Mask R-CNN model trained using the expanded dataset.

The results reveal that the loss exponentially decreases with increasing number of training epochs and that the loss reduction rate begins to plateau at 150 epochs. A loss value of approximately 0 usually indicates that the model's predictions are very close to the true values, suggesting that the model is nearing its optimal performance and the model parameters are well calibrated. The Mask R-CNN typically analyzes data much more quickly (10 s) than manual analysis. It can complete the preprocessing, peak identification, and qualitative analysis of coal-based liquids' chromatograms in just a few seconds to minutes, whereas manual analysis may take several hours or even days, depending on the complexity of the sample and the experience of the analyst. When dealing with large amounts of data, the Mask R-CNN can reduce human errors and provide more consistent and reliable results.

The performance of the Mask R-CNN model in the task of identifying the GC-GC chromatograms of DCL oils and classifying their different compositions is excellent. The model achieves an accuracy of 93.71%, indicating that it gives correct predictions in most cases and has good overall classification capabilities. The precision rate of 96.99% reflects the high reliability of the model in predicting positive classes and shows that there are very few cases in which a negative class is mistakenly predicted to be positive. The recall is 96.27% shows that the model is able to identify the actual positive class samples well, with very few omissions, and has a high degree of coverage of the actual positive classes. The F1 value, which is the harmonic mean of the precision rate and the recall, has a value of 0.95 indicating that both precision and recall are at a high level and relatively balanced, and the model performs well in terms of precision and completeness. Additionally, IoU is used to evaluate the localization accuracy of target detection by calculating the ratio of the overlap between the real target frame and the actual detected target frame. The IoU value of 0.93 indicates that the overlap between the detected target frames and the real target frames is better higher, and the localization is more accurate.

Therefore, the integration of GC × GC chromatograms with DL-Mask R-CNN techniques offers a more efficient and rapid approach to coal-based liquids' compound analysis. The proposed method not only facilitates the visual comparison of chromatograms but also enables the differentiation of the chemical compositions of coal-based liquids.

3.2.5. Practical Applicability of the Mask R-CNN

Although the Mask R-CNN effectively recognizes patterns in the chromatograms of DCL oil, it is still suitable for recognizing patterns in the chromatograms of other oils, such as coal tar and jet fuel. Because of its power and flexibility in segmenting images, the Mask R-CNN can be applied to recognize patterns in the chromatograms of other oils, especially for accurately recognizing and segmenting spectral features of oils. However, the Mask R-CNN must be appropriately adjusted and optimized to adapt to the spectral characteristics of different oils. For example, the network structure, loss function, and training strategy may need to be adjusted to improve the model's accuracy in recognizing the spectral features of specific oils. Transfer learning (TL) is a ML technique that allows a model to utilize knowledge learned on a related task on a new task. This approach is particularly suitable for situations where the amount of data is limited, by pre-training the model on a large-scale dataset and then applying it to a small-scale dataset for fine-tuning, thus improving the model's performance on the new task [36,37]. In the image recognition of DCL oil, acquiring a large number of GC × GC chromatograms can be very difficult and expensive. Employing TL to extract the classification features of DCL oils can reduce the dependence on the amount of data of chromatograms of other coal-based oils by utilizing the generalized features learned by the pre-trained model in the chromatograms of DCL oils, so that a better performance in classifying the substances in other coal-based oils, such as coal tar, can be achieved in spite of the limited amount of data. TL can significantly improve the performance of the model.

4. Conclusions

1. GC × GC is a highly efficient technique for analyzing complex mixtures that can provide more detailed information on molecular composition by separating compounds. Thirty-six DCL oils were qualitatively analyzed using GC × GC–MS. The oil components was classified into aliphatic alkanes and cycloalkanes, mono and polycyclic aromatic compounds, O-, N- and S-containing compounds. The chromatograms were segmented into six distinct regions corresponding to compound classes, providing a foundation for the subsequent pattern recognition step.
2. An analytical method was proposed for comprehensively characterizing the spots in the GC × GC chromatograms of DCL oils. Mask R-CNN, as a target detection and segmentation model, can effectively recognize and classify different constituents in DCL oil GC × GC chromatograms. This method is effective for visually comparing chromatograms. It utilizes the distributions of the spots in chromatograms and the Mask R-CNN to quickly segment GC × GC chromatograms into regions representing different compounds. This process automatically qualitatively classifies the compounds based on the spots in their corresponding chromatograms. The primary advantage of the method is its ability to efficiently process multiple chromatograms in batches, substantially accelerating the overall analysis and shortening manual analysis, thereby substantially enhancing efficiency;
3. The Mask R-CNN is particularly useful for accurately and rapidly analyzing the chemical compositions of multiple coal-based liquids, which is vital for further processing and utilization of coal-based liquids.

Author Contributions: H.-H.F., X.-L.W.: Formal analysis, investigation, Methodology; Writing—original draft; J.F.: Conceptualization, Resources, Supervision; W.-Y.L.: Funding acquisition, Writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work was financially supported by the Science and Technology Innovation Project of China Shenhua Coal to Liquid and Chemical Co., Ltd. (contract no. [2022]101), National Natural Science Foundation of China (grant number: 22038008), and Program of Beijing Huairou Laboratory (ZD2023012A).

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DCL	direct coal liquefaction
1D	one-dimensional
2D	two-dimensional
3D	three-dimensional
GC × GC	comprehensive two-dimensional gas chromatography
DL	deep learning
ML	machine learning
MS	mass spectrometry
FID	flame ionization detector
GC × GC-MS	comprehensive two-dimensional gas chromatography- mass spectrometry
Mask R-CNN	mask region-based convolutional neural network
LTS	long-term support
ROI	region of interest
FPN	feature pyramid network
RPN	region proposal network

THN	1,2,3,4-tetrahydronaphthalene
NIST	national institute of standards and technology
IoU	Intersection over Union
TL	Transfer learning

Appendix A

Mask R-CNN is an innovative neural network architecture designed for precise instance segmentation within images. Its structure primarily comprises a pre-trained CNN, such as ResNet, which serves as the base layer for feature extraction from input images. Built upon this foundation, a Feature Pyramid Network (FPN) generates a multi-scale feature pyramid, enhancing the model's capability to detect objects across various sizes. At each level of this pyramid, a Region Proposal Network (RPN) identifies potential regions of interest (ROIs) specific to objects within the image.

A key innovation of Mask R-CNN is the introduction of ROI Align, which ensures the preservation of spatial location information with enhanced accuracy. For each proposed ROI, the network predicts the object category and refines the bounding box coordinates. The Mask Branch, a distinctive feature of Mask R-CNN, dedicates a separate pathway for predicting detailed segmentation masks for each ROI, facilitating instance-level segmentation.

The framework utilizes a multi-task loss function that accounts for classification, bounding box regression, and mask prediction losses, optimizing all tasks simultaneously. During training, the network learns to map image pixels to class labels, bounding boxes, and segmentation masks. In the inference phase, it can accurately segment instances in new images. The integration of these components empowers Mask R-CNN to achieve not only precise segmentation of individual instances but also high efficiency in object detection, making it a powerful tool for image analysis tasks.

References

1. Li, W.Y.; Wang, X.L.; Fan, H.H.; Fan, H.X.; Feng, J. Predicting the fuel performance of coal-based liquids using the ML-QSPR method. *J. Coal Sci. Eng. China* **2024**, *49*, 1098–1110.
2. Liu, Z.Y.; Phillips, J.B. Comprehensive Two-Dimensional Gas Chromatography using an On-Column Thermal Modulator Interface. *J. Chromatogr. Sci.* **1991**, *29*, 227–231. [CrossRef]
3. Pollo, B.J.; Alexandrino, G.L.; Augusto, F.; Hantao, L.W. The impact of comprehensive two-dimensional gas chromatography on oil & gas analysis: Recent advances and applications in petroleum industry. *Trends Analyt. Chem.* **2018**, *105*, 202–217.
4. Klee, M.S.; Cochran, J.; Merrick, M.; Blumberg, L.M. Evaluation of conditions of comprehensive two-dimensional gas chromatography that yield a near-theoretical maximum in peak capacity gain. *J. Chromatogr. A* **2015**, *1383*, 151–159. [CrossRef]
5. Lee, A.L.; Bartle, K.D.; Lewis, A.C. A Model of Peak Amplitude Enhancement in Orthogonal Two-Dimensional Gas Chromatography. *Anal. Chem.* **2001**, *73*, 1330–1335. [CrossRef]
6. Khalturin, A.A.; Parfenchik, K.D.; Shpenst, V.A. Features of Oil Spills Monitoring on the Water Surface by the Russian Federation in the Arctic Region. *J. Mar. Sci. Eng.* **2023**, *11*, 111. [CrossRef]
7. Kallio, M.; Hyötyläinen, T. Simple calibration procedure for comprehensive two-dimensional gas chromatography. *J. Chromatogr. A* **2008**, *1200*, 264–267. [CrossRef]
8. Shellie, R.; Marriott, P.; Morrison, P. Concepts and Preliminary Observations on the Triple-Dimensional Analysis of Complex Volatile Samples by Using GC×GC−TOFMS. *Anal. Chem.* **2001**, *73*, 1336–1344. [CrossRef]
9. Yuan, S.Y.; Li, H.J.; Liu, Z.Q.; Wang, Y.T.; Li, W.; Zhang, X.W.; Liu, G.Z. Measurement of non-hindered and hindered phenolic species in aviation fuels via tandem-SPE with comprehensive GC×GC–MS/FID. *Fuel* **2020**, *287*, 119561. [CrossRef]
10. Trinklein, T.J.; Cain, C.N.; Ochoa, G.S.; Schöneich, S.; Mikaliunaite, L.; Synovec, R.E. Recent Advances in GC×GC and Chemometrics to Address Emerging Challenges in Nontargeted Analysis. *Anal. Chem.* **2023**, *95*, 264–286. [CrossRef]
11. Furbo, S.; Hansen, A.B.; Skov, T.; Christensen, J.H. Pixel-Based Analysis of Comprehensive Two-Dimensional Gas Chromatograms (Color Plots) of Petroleum: A Tutorial. *Anal. Chem.* **2014**, *86*, 7160–7170. [CrossRef] [PubMed]
12. Sudol, P.E.; Pierce, K.M.; Prebihalo, S.E.; Skogerboe, K.J.; Wright, B.W.; Synovec, R.E. Development of gas chromatographic pattern recognition and classification tools for compliance and forensic analyses of fuels: A review. *Anal. Chim. Acta* **2020**, *1132*, 157–186. [CrossRef] [PubMed]

13. Jennerwein, M.K.; Eschner, M.; Gröger, T.; Wilharm, T.; Zimmermann, R. Complete Group-Type Quantification of Petroleum Middle Distillates Based on Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry (GC×GC-TOFMS) and Visual Basic Scripting. *Energy Fuels* **2014**, *28*, 5670–5681. [CrossRef]
14. Lissitsyna, K.; Huertas, S.; Quintero, L.C.; Polo, L.M. PIONA analysis of kerosene by comprehensive two-dimensional gas chromatography coupled to time of flight mass spectrometry. *Fuel* **2014**, *116*, 716–722. [CrossRef]
15. Rathsack, P.; Otto, M. Classification of chemical compound classes in slow pyrolysis liquids from brown coal using comprehensive gas-chromatography mass-spectrometry. *Fuel* **2014**, *116*, 841–849. [CrossRef]
16. Fan, Y.J.; Yu, C.X.; Lu, H.M.; Chen, Y.; Hu, B.B.; Zhang, X.R.; Su, J.E.; Zhang, Z.M. Deep learning-based method for automatic resolution of gas chromatography-mass spectrometry data from complex samples. *J. Chromatogr. A* **2023**, *1690*, 463768. [CrossRef]
17. Kehimkar, B.; Hoggard, J.C.; Marney, L.C.; Billingsley, M.C.; Fraga, C.G.; Bruno, T.J.; Synovec, R.E. Correlation of rocket propulsion fuel properties with chemical composition using comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry followed by partial least squares regression analysis. *J. Chromatogr. A* **2014**, *1327*, 132–140. [CrossRef]
18. Liu, J.W.; Ahmad, F.; Zhang, Q.; Liang, L.T.; Xiang, X.N. Interactive tools to assist convenient group-type identification and comparison of low-temperature coal tar using GC×GC–MS. *Fuel* **2020**, *278*, 118314. [CrossRef]
19. Jiang, B.D.; An, X.Y.; Xu, S.F.; Chen, Z.L. Intelligent Image Semantic Segmentation: A Review Through Deep Learning Techniques for Remote Sensing Image Analysis. *J. Indian Soc. Remote. Sens.* **2023**, *51*, 1865–1878. [CrossRef]
20. Archana, R.; Jeevaraj, P.S.E. Deep learning models for digital image processing: A review. *Artif. Intell. Rev.* **2024**, *57*, 11. [CrossRef]
21. Choudhary, K.; DeCost, B.; Chen, C.; Jain, A.; Tavazza, F.; Cohn, R.; Park, C.W.; Choudhary, A.; Agrawal, A.; Billinge, S.J.L.; et al. Recent advances and applications of deep learning methods in materials science. *NPJ Comput. Mater.* **2022**, *8*, 59. [CrossRef]
22. Babu, B.R.; Kiran, S. An Analysis of Deep Learning-Based Image Segmentation Techniques. In *Soft Computing for Security Applications*; Ranganathan, G., El Allioui, Y., Piramuthu, S., Eds.; Springer Nature Singapore: Singapore, 2023; pp. 725–737.
23. Kaur, R.; Singh, S. A comprehensive review of object detection with deep learning. *Digit. Signal Process.* **2023**, *132*, 103812. [CrossRef]
24. Kan, A. Machine learning applications in cell image analysis. *Immunol. Cell Biol.* **2017**, *95*, 525–530. [CrossRef]
25. He, K.M.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
26. He, K.M.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 386–397. [CrossRef] [PubMed]
27. Shafizadeh, A.; Shahbeik, H.; Rafiee, S.; Fardi, Z.; Karimi, K.; Peng, W.; Chen, X.; Tabatabaei, M.; Aghbashlo, M. Machine learning-enabled analysis of product distribution and composition in biomass-coal co-pyrolysis. *Fuel* **2024**, *355*, 129464. [CrossRef]
28. Hollingsworth, B.V.; Reichenbach, S.E.; Tao, Q.; Visvanathan, A. Comparative visualization for comprehensive two-dimensional gas chromatography. *J. Chromatogr. A* **2006**, *1105*, 51–58. [CrossRef] [PubMed]
29. Vendeuvre, C.; Ruiz-Guerrero, R.; Bertoncini, F.; Duval, L.; Thiébaut, D.; Hennion, M.-C. Characterisation of middle-distillates by comprehensive two-dimensional gas chromatography (GC×GC): A powerful alternative for performing various standard analysis of middle-distillates. *J. Chromatogr. A* **2005**, *1086*, 21–28. [CrossRef] [PubMed]
30. Reichenbach, S.E.; Kottapalli, V.; Ni, M.; Visvanathan, A. Computer language for identifying chemicals with comprehensive two-dimensional gas chromatography and mass spectrometry. *J. Chromatogr. A* **2005**, *1071*, 263–269. [CrossRef] [PubMed]
31. Schmarr, H.-G.; Bernhardt, J. Profiling analysis of volatile compounds from fruits using comprehensive two-dimensional gas chromatography and image processing techniques. *J. Chromatogr. A* **2010**, *1217*, 565–574. [CrossRef]
32. Van Stee, L.L.P.; Brinkman, U.A.T. Peak detection methods for GC × GC: An overview. *Trends Anal. Chem.* **2016**, *83*, 1–13. [CrossRef]
33. Asnin, L.D. Peak measurement and calibration in chromatographic analysis. *Trends Anal. Chem.* **2016**, *81*, 51–62. [CrossRef]
34. Matyushin, D.D.; Sholokhova, A.Y.; Buryak, A.K. Deep Learning Driven GC-MS Library Search and Its Application for Metabolomics. *Anal. Chem.* **2020**, *92*, 11818–11825. [CrossRef] [PubMed]
35. Ahmed, S.F.; Alam, M.S.B.; Hassan, M.; Rozbu, M.R.; Ishtiak, T.; Rafa, N.; Mofijur, M.; Shawkat Ali, A.B.M.; Gandomi, A.H. Deep learning modelling techniques: Current progress, applications, advantages, and challenges. *Artif. Intell. Rev.* **2023**, *56*, 13521–13617. [CrossRef]
36. Pan, S.J.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [CrossRef]
37. Tan, C.; Sun, F.; Kong, T.; Zhang, W.; Yang, C.; Liu, C. A Survey on Deep Transfer Learning. In Proceedings of the 27th International Conference on Artificial Neural Networks, Rhodes, Greece, 4–7 October 2018.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.