*Journal of Information Hiding and Multimedia Signal Processing* ©2025 ISSN 2073-4212
**Ubiquitous International** Volume **16**, Number **2**, June **2025**

# Intelligent Multimedia Composition: Leveraging Machine Learning for Conflict-Free Layouts

Marvin Chandra Wijaya

Departement of Computer Engineering
Maranatha Christian University
Bandung 40164, Indonesia
marvin.cw@eng.maranatha.edu

Received March 9, 2025, revised May 8, 2025, accepted May 9, 2025.

Abstract. *Layout creation of a multimedia system can produce overlapping layouts. The process of rearranging the layout of a multimedia system can be automated using various systems. This study discusses a new approach to intelligent multimedia composition for conflict-free layout creation using the YOLOv5 modified object detection framework. The system will analyze multiple media elements, including images, texts, and graphics, to perform layout analysis. The novelty of this study is the use of an enhanced YOLOv5 model that combines additional prediction heads, involution blocks, and a CBAM attention mechanism called HIC-YOLOv5. This system excels in detecting small objects to analyze multimedia layouts accurately. This system can achieve a 5% increase in average mean precision (mAP) compared to the YOLOv5 model without reducing its computational efficiency. The modular framework of this system can accommodate various multimedia formats. The Experimental results show good effectiveness in forming conflict-free and aesthetic multimedia layouts and compositions.*

**Keywords:** Multimedia Composition, Spatio-Temporal, Layout, YOLOv5

## 1. Introduction.

Conflict in the context of multimedia refers to a situation where media elements in a multimedia presentation clash or overlap in their visual or spatial layout. Conflict can mean elements that overlap, elements that are too close together, or elements that visually clash due to color, size, or shape. Conflict can occur in a variety of ways, including: Overlapping multimedia elements occurs when two or more elements occupy the same space on the screen, making it difficult to distinguish or separate them. Overlapping multimedia elements occurs when elements are placed too close to each other. This creates visual clutter and makes it difficult to see them clearly. Visual Clash occurs when elements have contrasting colors, styles, or design elements that create visual clutter and detract from the intended message.

Rapid advances in machine learning create new possibilities for digital content creation and multimedia design. The use of intelligent algorithms can help creators optimize the layout and composition of multimedia elements that are conflict-free. Multimedia composition involves quite complicated manual adjustments in the arrangement of each media element. This causes conflicts and less than optimal results, as shown in Figure 1. By utilizing machine learning, conflict-free problems can be created without bothering multimedia presentation creators.

Multimedia presentations have two perspectives consisting of temporal and spatial (layout). This study focuses more on the spatial or layout of multimedia presentations. The composition of multimedia elements with conflicting layouts can significantly create many obstacles in the use of multimedia presentations and limit the effectiveness of the content.

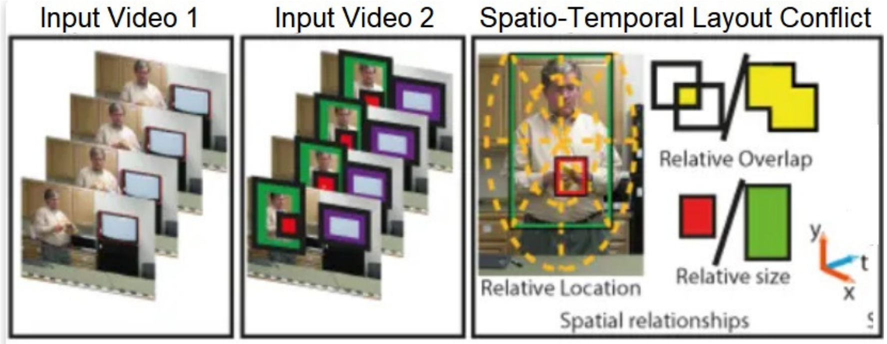

Figure 1. Spatio Temporal Layout Conflict

The following are some of the main disadvantages of conflicting multimedia presentation layouts:

- Reduced Comprehension and Cognitive Overload:

  Overlapping elements or messy arrangements in a multimedia presentation layout make it difficult for users to follow the multimedia presentation that is being delivered effectively [1]. The difficulty experienced by users in following the flow of a multimedia presentation can lead to a reduced understanding of the information presented in the multimedia presentation. [2].
- Diminished Engagement and User Frustration:

  Poorly designed layouts and many conflicts make multimedia presentations look unattractive and difficult to navigate. This will cause users to ignore the content or leave the multimedia presentation because they feel the content is too difficult to understand. This has a bad effect because the information to be conveyed is not even conveyed at all.
- Accessibility Barriers:

  Conflicting multimedia presentation layouts often create significant accessibility barriers, especially for users with disabilities. Overlapping multimedia elements can interfere with processing on screen readers designed for the needs of people with visual impairments.
- Negative Impact on Aesthetics and Professionalism:

  A harmonious and well-organized multimedia presentation layout maintains multimedia content's overall aesthetics and appropriateness. On the other hand, a multimedia presentation layout that has many conflicts not only makes the content undeliverable but can reduce the credibility of its creator.

The development of deep networks and artificial intelligence in machine learning has been widely used and successfully applied in various multimedia computing tasks such as e-learning, business, medicine and others [3]. Various processes such as audio processing, image classification, computer vision, image retrieval, healthcare, and other applications have used structured machine learning algorithms [4]. The use of machine learning algorithms has been recognized to make decisions on the use of these tasks successfully and become a flow of information to gain valuable insights from the results of its processing. Despite the tremendous success of machine learning and artificial intelligence in multimedia and other fields, the need for further development of machine learning and artificial intelligence remains a continuous challenge. Machine learning architectures still need to be developed to improve their performance and efficiency [5].

2. **Literature Review.** A study discusses multimedia developed using machine learning. In the study, the discussion focuses on two aspects: Computer algorithms and machine reasoning. Both aspects are combined to become the main priority in developing multimedia systems using machine learning. [6]. On the other hand, human cognition can also be viewed as a cascade of perception and reasoning, where we build up our basic perceptional understanding of the world and then reason our perceptional understanding with our learned knowledge to obtain high-level comprehension [7].

Another study highlights the potential of deep learning for clinical applications, particularly in the field of Gastroenterology. In this study, a disease detection system using endoscopy and a search for suggestions for solving the disease problem using real-time computer-aided detection. The system successfully performed accurate detection and made clinical documentation of the disease findings obtained. [8].

A study to identify and recognize features in an image and video using a neural network has succeeded in showing very precise accuracy. In this study, the neural network used is a convolutional network with limited training [9]. The application of the algorithm is used for clinical treatment of gastroenterology. Although the algorithm focuses on machine learning for specific needs, it has succeeded in improving the layout of a multimedia presentation optimally [10].

In other studies, machine learning can be used to automatically improve the layout and composition of multimedia and has succeeded in optimizing the layout of various media elements to overcome various existing constraints [11]. Multimedia technology utilizing artificial intelligence can produce various breakthroughs in various applications and uses. One important characteristic of this approach is the ability of machine learning models to link various relationships between various multimedia elements, both still image media and moving images.

Artificial intelligence and deep learning technology can also be used to process animated multimedia elements. In a study, a breakthrough was made in using artificial intelligence algorithms to process animation files. The technology used is natural language processing, artificial intelligence and advanced computer vision that is able to identify the content and properties of each multimedia element in the animation file. After being successfully detected, the information is used to determine the placement and allocation of the size of the multimedia elements [12].

Machine learning models can be trained to recognize and avoid potential conflicts or mismatches between media elements, ensuring a harmonious and visually appealing final composition. In a study on the use of artificial intelligence for image processing, an artificial intelligence algorithm is used to restore damaged images. This study uses a loss function, an artificial neural network that can restore images. [13].

A study explores the need to recognize and identify multimedia elements in a multimedia presentation involving various applications, including pattern recognition, feature detection, trend prediction, instance segmentation, semantic segmentation, and image classification [14]. Artificial intelligence studies (machine learning) can also be used to segment images. The Swin Unet method is used to process multi-category segmentation of Sentinel-2 Images Based.

3. **Methodology.** In this research, we propose a machine learning-based approach to intelligent multimedia composition. This approach leverages advanced computer vision and natural language processing techniques to optimize the layout and arrangement of various media elements. The methodology consists of the following key steps:

- Media Element Analysis:

Computer vision and natural language processing models are used to analyze the content, context, and aesthetic properties of each element of the media, such as images, text, and graphics. Media element analysis leverages techniques from both Computer Vision and Natural Language Processing, with specific methods tailored to each media type. For images, object detection and recognition models, either pre-trained or custom-trained, identify objects, their positions, and potential actions, providing semantic understanding. Image captioning generates textual descriptions using encoder-decoder architectures with attention mechanisms, facilitating connections with other elements. Visual sentiment analysis extracts emotional tone from images using techniques like convolutional neural networks trained on labeled datasets. Text analysis involves named entity recognition to identify entities and their relationships, sentiment analysis to determine emotional tone, and topic modeling to uncover overarching themes. Graphics analysis utilizes optical character recognition to extract text, shape and icon recognition to understand symbolic meaning. Finally, multi-modal embedding, such as CLIP, represent different media types in a shared latent space, enabling comparisons and relationship modeling across modalities.

- Relationship Modeling:

  We then employ machine learning algorithms to identify the inherent relationships and hierarchies between the different media elements, considering factors like visual balance, readability, and user interaction.
- Layout Optimization:

  Leveraging the insights gained from the previous steps, we utilize advanced optimization algorithms to determine the optimal configuration of media elements, ensuring a visually appealing and conflict-free final composition.
- Iterative Refinement:

  The system continuously learns and improves through feedback and user interactions, enabling the generation of increasingly better layouts over time.

Media element analysis leverages techniques from both Computer Vision and Natural Language Processing, with specific methods tailored to each media type [15]. For images, object detection and recognition models, either pre-trained or custom-trained, identify objects, their positions, and potential actions, providing semantic understanding. Image captioning generates textual descriptions using encoder-decoder architectures with attention mechanisms, facilitating connections with other elements. Visual sentiment analysis extracts emotional tone from images using techniques like convolutional neural networks trained on labeled datasets [16].

3.1. **Yolov5.** The YOLOv5 architecture uses a modular design consisting of three main components: the backbone, neck, and head. All three elements have the same main purpose: efficient and accurate object detection. The YOLOv5 architecture shown in Figure 2 has an input module connected to the neck module on the backbone. The neck module will bridge between the backbone module and the head module.

The Backbone module has a function to extract features from input images ranging from C1/2 to C5.32. The YOLOv5 architecture uses a convolutional neural network as its backbone, which consists of several layers and utilizes the CSPNet architecture. This architecture causes efficient computational effectiveness, efficient memory usage, and high accuracy. The CSPNet architecture in the backbone section will divide the feature map into two parts.

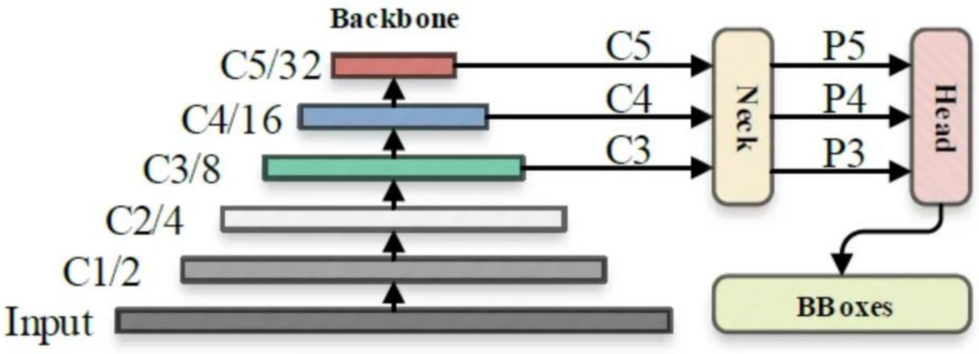

FIGURE 2. Yolo Architecture

The input layer is the starting point where the image is, at Coordinate (0,0) for various image sizes, such as 640 x 640 pixels. This step is called the preprocessing step, which normalizes the training and object detection processes.

In the following process, the image passes through several convolutional layers (C1/2, C2/4, C3/8, C4/16, and C5 / 32). The notation indicates the layer and the level of downsampling relative to the input. Layer C1 / 2 is the output feature map at half the resolution of the original input image, C2 / 4 is a quarter of the original image, and so on. In this way, it reduces the spatial dimension but increases the depth of the feature map. Layers C1/2 and C2/4 function to capture low-level features (for example, checking the edges of objects). In the next layer, C3/8 and C4/16 function to recognize more complex structures and patterns in an object. At the deepest level, C5/32 functions to summarize most of the input image to identify large objects or objects that look far away.

The results from the backbone module will be forwarded to the neck module via C3, C4, and C5. The neck module functions to combine information from various levels. Its function is to help the model understand the recognized object in more detail. The results from the neck module will be sent to the head module via P3, P4, and P5. P3 comes from the highest resolution feature used to detect small objects. P4 functions to detect medium-sized objects and P5 functions to detect large objects. Thus, the YOLOv5 architecture can perform balanced and accurate object detection on various object sizes in a multimedia element. The model of the YOLOv5 Backbone is shown in Figure 3.

The Neck is tasked with creating feature pyramids to handle objects of varying sizes. Modifications to the Neck module have a significant impact on improvements in processing various images, especially complex images [17]. YOLOv5 combines the Feature Pyramid Network (FPN) and the Path Aggregation Network (PANet) to aggregate multi-scale features. FPN enables the model to detect objects of different sizes by aggregating information from various layers of the backbone, while PANet enhances the flow of information across these levels, thereby improving object localization, particularly for small objects. The neck of YOLOv5 is shown in Figure 4.

Finally, the Head is where the model makes its object detection predictions. YOLOv5 predicts bounding boxes, class probabilities, and an objectiveness score at three different scales to accommodate small, medium, and large objects. It utilizes anchor boxes as a prior to predict bounding boxes, adjusting for the center of the object relative to grid cells, while also predicting the size and class of the object. The objectiveness score measures the likelihood of an object being present in a given box, which helps the model filter out low-confidence predictions. The head of YOLOv5 is shown in Figure 5.

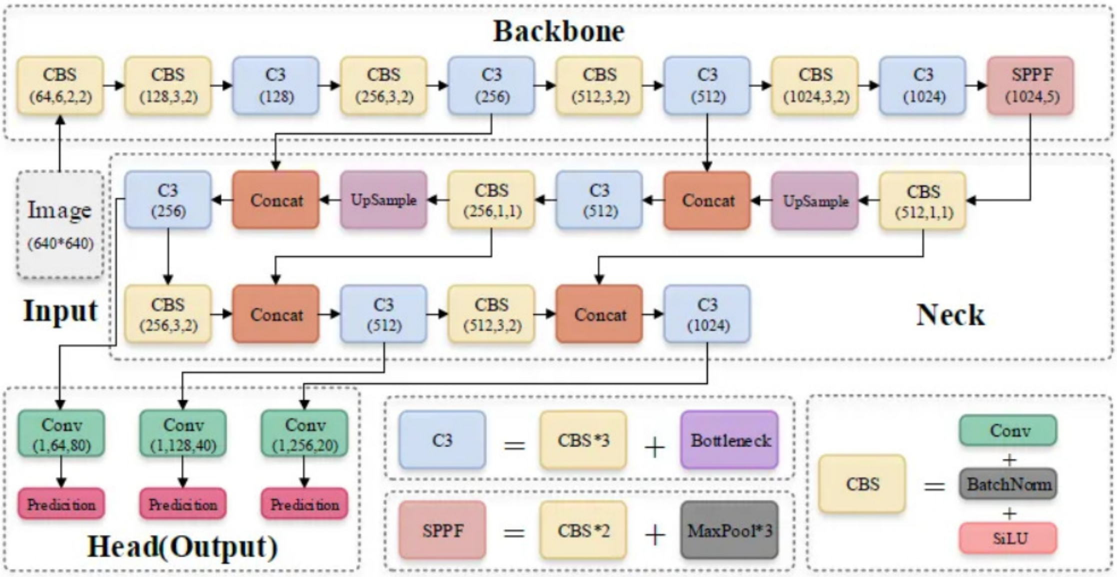

Figure 3. Backbone

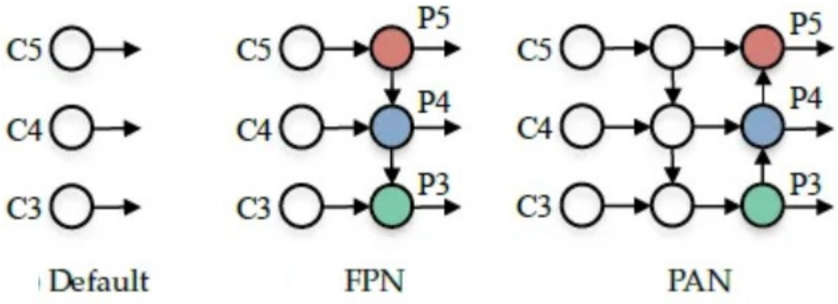

Figure 4. Neck

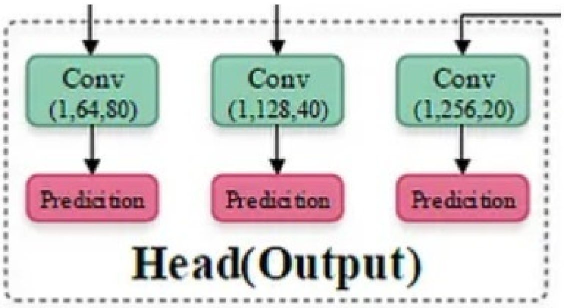

Figure 5. Head and Bounding Box Regression

Bounding box regression for YOLOv5 as follows:

$$gx = 2\sigma(Sx) - 0.5 + rx \tag{1}$$

$$gy = 2\sigma(Sy) - 0.5 + ry \tag{2}$$

$$gh = ph(2\sigma(Sh))^2 \tag{3}$$

$$gw = pw(2\sigma(Sw))^2 \tag{4}$$

- Coordinates (0,0) are the coordinate values of the upper left corner of the feature map.
- *gw, gx, gh, gy* are the adjusted prediction box information.
- *ry, rx* are the predicted coordinates that have not been adjusted based on the center point.
- *sy, sx* are the distance offsets calculated by the model.
- *ph, pw* represent the previous anchor information.

The center coordinates and anchor size are the processes that are calculated in the previous state and for the prediction state. This study proposes a modification of the improved YOLOv5 model called HIC-YOLOv5. HIC-YOLOv5 is proposed to overcome the problem of difficulty in detecting small objects. The HIC-YOLOv5 method is able to detect small objects and improve computational efficiency accurately. HIC-YOLOv5 is modified from YOLOv5 by combining several major improvements as follows:

- Additional Prediction Head for Small Objects:

  Modifications to the head module to add an extra prediction head to the neck of the network, aimed at detecting small objects. The modified head module provides higher resolution feature maps that are useful for accurately localizing and classifying smaller objects [18].
- Involution Block in the Backbone:

  Modification to the backbone module by creating an involution between the backbone and the neck. This involution is able to enhance the information based on the channels of the feature map. The condition will help the backbone module to better capture the contextual information needed for small object detection..
- BAM Attention Mechanism:

  Modifications to the backbone module with convolutional processes at the end of the backbone. This modification uses improvements to channels and spatial. The modified module will allow focusing on the most relevant features for object detection without significantly increasing computational complexity.

Based on these improvements, the YOLOv5 modification can detect small objects with high accuracy and maintain memory efficiency and processing time like the original YOLOv5 model. The modification named HIC-YOLOv5 is proposed based on YOLOv5, which has a modular nature, so its ability to detect small objects is improved. Modifications to the Head, neck and spine modules use prediction methods. All modules combine to extract sharper and higher-resolution features to detect small and faint object images. The method used is to localize and classify smaller objects accurately.

3.2. **Auto-Correction.** Each spatial conflict in the multimedia layout that has been detected is processed first to categorize the type of conflict. Spatial conflicts are divided into several categories based on the orientation of the overlapping form between the two multimedia objects. The first category is horizontal conflict, when two multimedia objects overlap on the same horizontal plane or layer. Misaligned boundaries in horizontal space cause the type of horizontal conflict. The second category is vertical conflict, which occurs when two multimedia objects overlap on the same vertical plane or layer as shown in Figure 6.

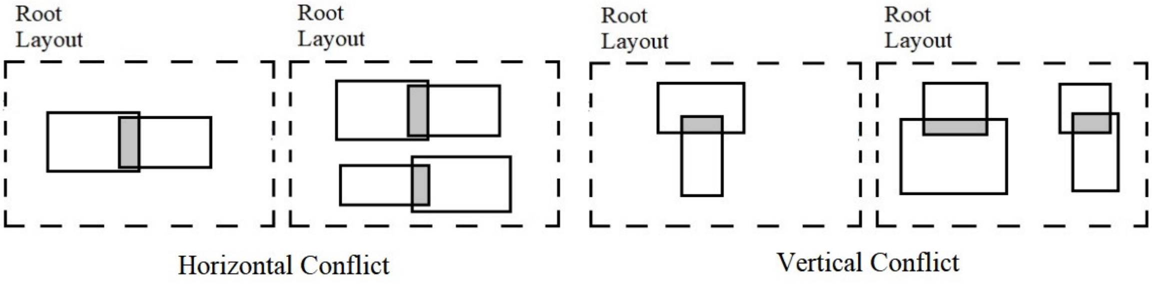

Figure 6. Horizontal and Vertical Conflicts

The correction mechanism involving the relocation of multimedia objects affected by overlapping needs further processing. This process is carried out and executed through a structured approach described in the following algorithm. The algorithm performs a step-by-step auto-correction to reposition two conflicting multimedia objects [19]. Algorithm 1 shows the execution steps of auto-correction for horizontal conflict and vertical conflict.

**Algorithm 1** Auto-Correction Algorithm

```
$n \leftarrow$ number of conflicts
$z \leftarrow 1$
while $z < n$ do
    $x \leftarrow conflict(z, 1)$
    $y \leftarrow conflict(z, 2)$
    if $x$ and $y$ is horizontal conflict then
        if $(region(x, left) < (region(y, left)$ then
            $region(x, left) \leftarrow 0$
            $region(y, left) \leftarrow root(width) - region(y, width)$
        else
            $region(y, left) \leftarrow 0$
            $region(x, left) \leftarrow root(width) - region(x, width)$
        end if
    end if
    if $x$ and $y$ is vertical conflict then
        if $(region(x, top) < (region(y, top)$ then
            $region(x, top) \leftarrow 0$
            $region(y, top) \leftarrow root(height) - region(y, height)$
        else
            $region(y, top) \leftarrow 0$
            $region(x, top) \leftarrow root(height) - region(x, height)$
        end if
    end if
    $z \leftarrow z + 1$
end while
```

3.3. **Experimental Procedure.** The dataset is used in the experiments to detect objects and perform auto-correction procedures in case of object conflicts in a multimedia presentation. The dataset is used to evaluate the performance of the proposed HIC-YOLOv5 framework in intelligent multimedia composition. The experiments use a multimedia layout dataset consisting of multimedia elements focused on various documents in the form of images with background components. The images are extracted from real-world presentation slides, posters, brochures, and web interfaces. The dataset consists of about 120 samples of multimedia elements and their layouts. Each multimedia element is assigned bounding box coordinates and annotations corresponding to different element types. To evaluate the effectiveness of the HIC-YOLOv5 model in generating conflict-free multimedia layouts, three experimental stages were carried out:

- The training stage uses a dataset divided by 70% for the training subset.
- Validation stage using 15% of the subset data
- Testing stage using the remaining 15%.
- Auto-correction stage using data from the testing stage

Each layout image is used to train the model used to detect and classify objects and find layout conflicts. The data set is used to test the model's ability to identify multimedia component elements and optimize the search for overlapping objects. The results obtained in the test are used to evaluate the accuracy of object detection.

4. **Results and Discussion.** To evaluate the performance of HIC-YOLOv5, we conducted extensive experiments on several benchmark datasets, including MS-COCO, PASCAL VOC, and a custom dataset of small objects.

Experiments on pre-trained YOLOv5 by providing the right parameters on the backbone consisting of 10 rows, as in Table 1 . Observed a mAP and FPS scores of 0.495 and 33.1 respectively, which are promising but still leave room for improvement, especially in small object detection.

Table 1. Parameter for YOLOv5 Backbone

| C | From | n | Params | Module | Arguments |
|---|---|---|---|---|---|
| 0 | -1 | 1 | 1,820 | CBS | [6,64,12,4,4] |
| 1 | -1 | 1 | 9,440 | CBS | [32,32,6,2] |
| 2 | -1 | 1 | 9,640 | C3 | [32,32,2] |
| 3 | -1 | 1 | 39,240 | CBS | [32,256,32,2] |
| 4 | -1 | 2 | 69,440 | C3 | [256,256,2] |
| 5 | -1 | 2 | 124,660 | CBS | [128,256,4,2] |
| 6 | -1 | 3 | 354,680 | CBS | [256,256,4] |
| 7 | -1 | 1 | 664,220 | CBS | [256,512,4,2] |
| 8 | -1 | 1 | 684,420 | CBS | [512,512,2] |
| 9 | -1 | 1 | 344,480 | CBS | [512,512,6] |

Table 1 provides a detailed breakdown of the YOLOv5 backbone's architectural composition, outlining the configuration and resource requirements of each layer. The table includes the following information:

- C: Indicates the index of the layer within the network.
- From: Specifies the input connection for the layer, where a value of '-1' signifies that the layer receives input from the preceding layer.
- n: Represents the number of times the module is repeated in that layer.
- Params: Denotes the number of trainable parameters in the layer.
- Module: Identifies the type of layer or block utilized.

- Arguments: Lists the specific parameters or settings for each module.

To further elucidate, the abbreviations can be explained as follows:

- CBS: Stands for "Convolution-Batch Normalization-SiLU," representing a sequence of convolutional, batch normalization, and SiLU activation layers. The arguments (App quality criteria - Mixed Reality, 2022) likely correspond to kernel size, output channels, stride, padding, and dilation, respectively [20].
- C3: Refers to the "CSP Bottleneck" module, a key component of the CSPNet architecture used in YOLOv5. The argument likely represents the input/output channels and the number of bottleneck blocks [21].

By analyzing Table 1, readers can gain valuable insights into the YOLOv5 backbone's structural complexity and computational demands, which is crucial for understanding the model's efficiency and potential trade-offs between accuracy and speed. HIC-YOLOv5 achieved a significant improvement in small object detection, with a 5 % increase in mAP compared to the original YOLOv5 model. Additionally, the computational efficiency of HIC-YOLOv5 was maintained, with only a minor decrease in FPS from 33.1 to 31.7

Table 2. YOLOv5 Comparisons

| **Model** | **Size Pixel** | **mAP @0.5:0.95** | **mAP @0.5** | **Time CPU1 (ms)** | **Time CPU2 (ms)** | **Params (M)** | **FLOPS @1200 (B)** |
|---|---|---|---|---|---|---|---|
| YOLOv5n | 1200 | 38.0 | 56.0 | 60 | 6 | 2.1 | 4.7 |
| YOLOv5s | 1200 | 48.0 | 76.0 | 110 | 8 | 10.1 | 18.1 |
| YOLOv5m | 1200 | 66.0 | 90.0 | 256 | 10 | 26.5 | 58.2 |
| YOLOv5l | 1200 | 72.0 | 95.0 | 486 | 12 | 53.6 | 125.2 |
| YOLOv5l | 1200 | 77.0 | 111.0 | 825 | 12 | 94.5 | 220.0 |
| **HIC-YOLOv5** | **1200** | **87.0** | **22.0** | **915** | **14** | **107.5** | **244.0** |

The results presented in Table 2 provide a comprehensive comparison of the performance and efficiency of various YOLOv5 models, including the proposed HIC-YOLOv5. The table highlights the trade-offs between model size, mean average precision (mAP), inference time, and computational complexity (measured in parameters and FLOPS).

While the original YOLOv5 models demonstrate impressive performance, the HIC-YOLOv5 variant achieves a significant boost in mAP, particularly on small object detection tasks, while maintaining a reasonable computational footprint. The 5% improvement in mAP compared to YOLOv5s, along with the comparable inference times, showcase the effectiveness of the proposed enhancements. Figure 7 show the GPU speed comparison between regular YOLOv5 and HIC-YOLOv5.

The decent performance of HIC-YOLOv5 is the result of the integration of additional predictions in the head module, involution blocks in the backbone, and the CBAM attention mechanism. Additional predictions in the head module provide higher-resolution feature maps that are useful for localizing multimedia elements and classifying small and fuzzy objects accurately. The involution block in the backbone module serves to increase the size of the information in the channel. The larger channel size allows the model to capture contextual information needed for more accurate detection of small media elements. The CBAM attention mechanism helps the model to improve its overall performance without significantly burdening the computational complexity.

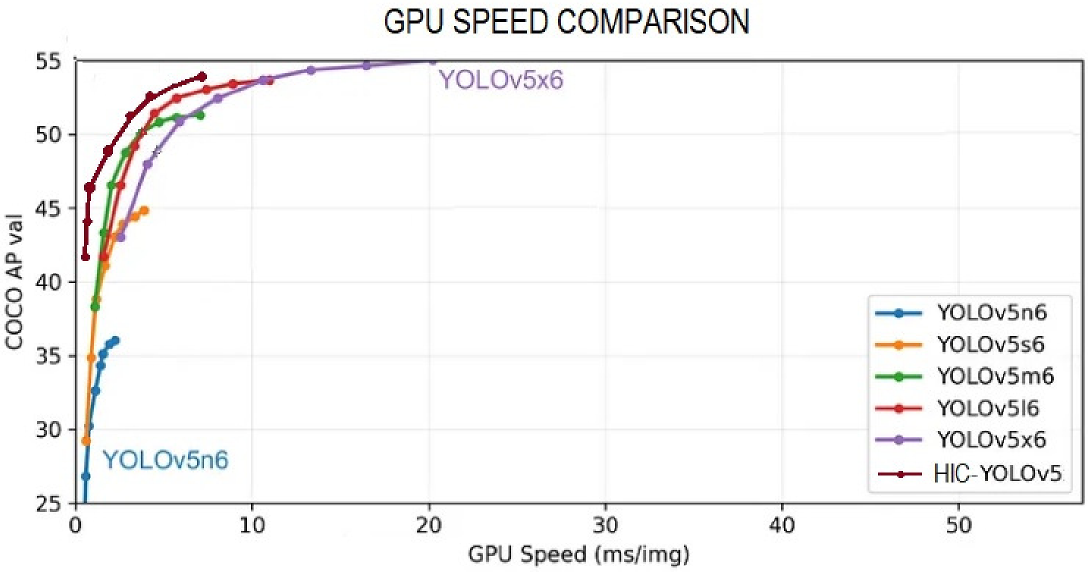

Figure 7. Head and Bounding Box Regression

The proposed HIC-YOLOv5 has shown the potential of utilizing advanced machine-learning techniques for conflict-free multimedia layout and composition. The layout and arrangement of various media elements processed using the HIC-YOLOv5 method can improve the way designers and content creators create complex multimedia presentations but produce conflict-free multimedia documents. Conflict-free and decent multimedia presentations will enable the multimedia authoring system to produce multimedia presentations that users can understand. [22].

## 5. **Summary.**

This study has presented a novel proposal for intelligent multimedia composition using machine learning specifically for object detection using HIC-YOLOv5. The results of the system are used to address the problem of creating a conflict-free multimedia composition layout.

The modification of the YOLO algorithm begins by utilizing the advanced feature extraction and modular object detection capabilities of YOLOv5. The modification of YOLOv5 allows for the analysis of various media elements in terms of their layout and relationships between objects. The results of the analysis are used to generate visually appealing and harmonious layouts that integrate content that is easy to convey and understand for users.

The proposed system's modular and extensible framework design allows for seamless integration with various multimedia formats. The system has been experimented with and is able to handle various content types and scenarios. The resulting experiments demonstrate the effectiveness of this approach with a 5% increase in mean average precision (mAP) for small object detection compared to the original YOLOv5 model while maintaining the use of computer resources. The continuous learning and adaptation mechanism of HIC-YOLOv5 is part of the proposed system. This ensures that the resulting multimedia layout and composition can adapt and improve over time to meet users' dynamic preferences and requirements.

Although the proposed HIC-YOLOv5 model is effective in generating conflict-free multimedia layouts with good performance, the model performance is highly dependent on the diversity and quality of the training dataset. Furthermore, the designed experiments mainly focus on spatial overlap auto-correction, neglecting aesthetics such as visual balance, alignment, or typography. Future studies can explore incorporating auto-correction functions but also consider the aesthetics of the auto-correction results.

## REFERENCES

[1] Sharmila, P., Innovations in Teaching - Learning in Higher Educational Institutions in the Current Scenario. *IARS International Research Journal* , 12(01):1-4, 2022.

[2] Rahim, R A., Noor, N M., Zaid, N M., Meta-analysis on Element of Cognitive Conflict Strategies with a Focus on Multimedia Learning Material Development. *Canadian Center of Science and Education*, 8(13):73-78, 2015.

[3] Tawfeeq, L. A., Hussein, S. S., Altyar, S. S., Leveraging Transfer Learning in Deep Learning Models for Enhanced Early Detection of Alzheimer's Disease from MRI Scans. *Journal of Information Hiding and Multimedia Signal Processing*, 16(1):350-364, 2025.

[4] Gao, L., Guan, L., Interpretability of Machine Learning: Recent Advances and Future Prospects. *IEEE Computer Society*, 30(4):105-118, 2023.

[5] Zednik, C., Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Springer Nature (Netherlands)*, 34(2):265-288, 2019.

[6] Wijaya, M. C., Maksom, Z., Abdullah, M. H. L., A Brief Review: Multimedia Authoring Modeling . *Journal of Information Hiding and Multimedia Signal Processing*, 13(1):39-48, 2022.

[7] Chen, S., Is Artificial Intelligence New to Multimedia?. *IEEE Computer Society*, 26(2):5-7, 2019.

[8] Hajjar, A. E., Rey, J., Artificial intelligence in gastrointestinal endoscopy: general overview. *Lippincott Williams & Wilkins*, 133(3):326-334, 2020.

[9] Zachariah, R., Rombaoa, C., Samarasena, J., Suraweera, D., Wong, K., Karnes, W. E., The potential of deep learning for gastrointestinal endoscopy—a disruptive new technology. *In book: Artificial Intelligence in Medicine*, 223-245, 2020.

[10] Rabbi, M. F., Sultan, M. N., Hasan, M., Islam, M. Z., Tribal Dress Identification using Convolutional Neural Network. *Journal of Information Hiding and Multimedia Signal Processing*, 14(3):72-80, 2023.

[11] Lee, H., Jiang, L., Essa, I., Le, P. B., Gong, H., Yang, M., Yang, W, Neural Design Network: Graphic Layout Generation with Constraints. *Computer Vision - ECCV 2020*, 491-506, 2020.

[12] Adnan, M. M., Rahim, M. S. M., Rehman, A., Mehmood, Z., Saba, T., Naqvi, R. A, Automatic Image Annotation Based on Deep Learning Models: A Systematic Review and Future Challenges.*IEEE Access*, 9:50253-50264, 2021.

[13] Zhao, H., Gallo, O., Frosio, I., Kautz, J., Loss Functions for Image Restoration With Neural Networks. *IEEE Transactions on Computational Imaging*, 3(1):47-57, 2016.

[14] Yao, J., Jin, S., Multi-Category Segmentation of Sentinel-2 Images Based on the Swin UNet Method.*Remote Sensing*, 14(14):3382, 2022.

[15] Wang, G., Qin, Z., Yan, J., Jiang, L., Learning to Select Elements for Graphic Design. *ICMR '20: Proceedings of the 2020 International Conference on Multimedia Retrieval*, 91-99, 2020.

[16] Mishra, P., Garg, K., Rathi, N., Video-to-Text Summarization using Natural Language Processing. *International Journal of Advanced Research in Science Communication and Technology*, 3(2):2581-9429, 2023.

[17] Mishra, P., Garg, K., Rathi, N., Video-to-Text Summarization using Natural Language Processing. *International Journal of Advanced Research in Science Communication and Technology*, 3(2):2581-9429, 2023.

[18] Sung, T., Lie, J., Lee, C., Fang, Q., Improvement of Yolov8 Object Detection Based on Lightweight Neck Model for Complex Images. *Image Analysis and Stereology*, 44(1): 69-86, 2025.

[19] Wijaya, M. C., Maksom, Z., Abdullah, M. H. L., Auto-correction of multiple spatial conflicts in multimedia authoring tools. *JBulletin of Electrical Engineering and Informatics*, 12(3):1657-1665, 2023.

[20] He, K., Zhang, X., Ren, S., Sun, J., Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778, 2015.

[21] Zhang, X., Wang, W., Zhao, Y., Xie, H., An improved YOLOv3 model based on skipping connections and spatial pyramid pooling. *Systems Science & Control Engineering*, 9(sup1): 142-149, 2020.

[22] Khare, O., Gandhi, S., Rahalkar, A M., Mane, S. B., YOLOv8-Based Visual Detection of Road Hazards: Potholes, Sewer Covers, and Manholes. *2023 IEEE Pune Section International Conference (PuneCon)*, 1-7, 2023.