HFM-YOLO: A Novel Lightweight and High-Speed Object Detection Model

Xinyi Gao, Minh Nguyen and Wei Qi Yan
Auckland University of Technology, 1010 New Zealand

ABSTRACT

In this book chapter, we introduce HFM-YOLO, a novel object detection model tailored for precise and efficient face mask detection. Based on the existing YOLOv8 framework, the model integrates the HGNetV2 backbone and RepConv layers while enhancing the object detection capabilities. Our evaluation using the Face Mask Detection dataset demonstrates HFM-YOLO's superior performance in precision, recall, and computational efficiency compared to the standard YOLO architectures. These results highlight its potential applicability in visual object detection.

Keywords: Object Detection, Deep Learning, Human Face Mask

INTRODUCTION

Efficient and accurate object detection models have been a primary focus of current research in the field of computer vision. The YOLO (You Only Look Once) series, particularly its most recent version YOLOv8 models, has made notable advancements in this field. Nevertheless, the increasing need for models that possess both exceptional precision and the ability to function effectively in the contexts with limited resources has resulted in the creation of HFM-YOLO (Human Face Mask-YOLO). This adaption of the YOLO framework is specifically designed for real-time detection of face masks.

The rise of worldwide health crises, notably the COVID-19 pandemic, has emphasised on the need for automated systems that can monitor adherence to health safety protocols, such as the use of face masks (Pollard et al., 2020). HFM-YOLO addresses this need by providing a compact but highly efficient technology to detect human faces and determine whether they are wearing a mask in different environments.

The focus of HFM-YOLO design is on replacing the traditional backbone network in YOLOv8 with HGNetV2 (Zhao et al., 2024), a smaller and more efficient network architecture. HGNetV2 is known for its superior combination of accuracy and efficiency. It can be further enhanced by using HGNetV2 as the backbone network, making it specialized for the tasks of detecting human face masks. This optimization requires strategically reducing network complexity in order to minimize computational requirements while maintaining the performance of visual object detection. The HFM-YOLO architecture incorporates RepConv (replaceable convolution) into its backbone network, effectively replacing traditional convolutional layers. Integrating RepConv into HGNetV2 significantly reduces the processing requirements of the network, thereby simplifying the model while maintaining its detection capabilities.

The use of RepConv in HFM-YOLO marks a significant shift towards lightweight model architectures in visual object detection. This approach not only increases processing speed, making it ideal for real-time applications. The approach also ensures that the deployment of the model is possible in resource-constrained scenarios such as mobile and embedded devices. Reducing model complexity does not reduce the ability of this model to identify complex face mask features, including a variety of mask types and occlusion directions.

Additionally, HFM-YOLO's design incorporates cutting-edge deep learning methods to ensure it remains robust in real-world applications. In terms of public health surveillance, the versatility and efficiency of HFM-YOLO make it a significant advance in visual object detection.

In summary, HFM-YOLO innovatively takes use of RepConv in HGNetV2. This model establishes a new benchmark among professional object detection models. This book chapter will further explore the structure, methodology, comprehensive performance evaluation, and its broader impact on real-time detection of HFM-YOLO model.

RELATED WORK

The field of visual object detection in computer vision has witnessed transformative changes (Xiao et al., 2021), in which the development of deep learning-based models is a key factor. This section provides an overview of key advances in the field, focusing on convolutional neural networks (CNN) (Yan, 2019), the evolution of YOLO family, and the innovative use of replaceable convolutions (RepConv) in object detection.

The emergence of CNN has revolutionized the field of computer vision. LeCun et al. (2015) introduced this concept, and computer vision began to enter the era of deep learning. The AlexNet model was subsequently proposed by Krizhevsky et al. (2012). This model further demonstrates the effectiveness of deep CNNs in image recognition tasks. After that, various architectures such as Vedaldi and Zisserman's VGG (Vedaldi & Zisserman, 2016) and He et al.'s ResNet (He et al., 2016) appeared one after another. Each model contributes to the efficiency and accuracy of CNN in object detection.

YOLO was proposed by Redmon et al. (2016) YOLO introduces a single-stage detection method. This method marks a major shift in target detection methods. In sharp contrast to previously dominant two-stage methods such as R-CNN by Girshick et al. (2014) The original YOLO model emphasized on speed but faced limitations in detection of small objects. Later versions, including YOLOv2 (Redmon & Farhadi, 2017) and YOLOv3 (Redmon & Farhadi, 2018), introduced improvements such as anchor boxes and multi-scale prediction, improving detection accuracy. Both YOLOv4 proposed by Bochkovskiy et al. (2020) and YOLOv5 (Jocher et al., 2022) of Ultralytics took use of the CSPDarknet53 advanced model. These models improve the efficiency and performance of visual object detection.

During the evolution of YOLO series, YOLOv6, YOLOv7 and YOLOv8 have promoted industrial applications. YOLOv6 (Li et al., 2022) combines processes such as EfficientRep, self-distillation and advanced quantification. It also provides a deployable network with customizable architecture, effectively balancing accuracy and speed. YOLOv7 (Wang et al., 2023) is an enhanced version of YOLOv5. YOLOv7 focuses on the training process and introduces strategies such as reparameterization modules and model scaling. YOLOv8 (Ju & Cai, 2023) further evolved from YOLOv5. At its core, it replaces the C3 module of its network backbone with C2f and adopts a decoupling process in Head. Together, these releases demonstrate significant advances in object detection performance and efficiency.

RepConv has attracted considerable interests in deep learning community. RepConv was firstly proposed by Soudy et al. (2023) RepConv is a convolution that is highly adaptable and can effectively replace traditional convolutional layers. Owing to its design, it can be easily replaced and modified within established CNN frameworks. By replacing traditional convolutions, neural networks that enhance specific goals can be achieved.

METHODOLOGY

HFM-YOLO is a specialized object detection model, which is designed to detect masks efficiently and accurately. The architecture of HFM-YOLO is based on YOLOv8 framework but with significant modifications. The modified model is more suitable for face mask detection. The model architecture integrates HGNetV2 as its backbone by replacing traditional convolutional layers with RepConv layers.

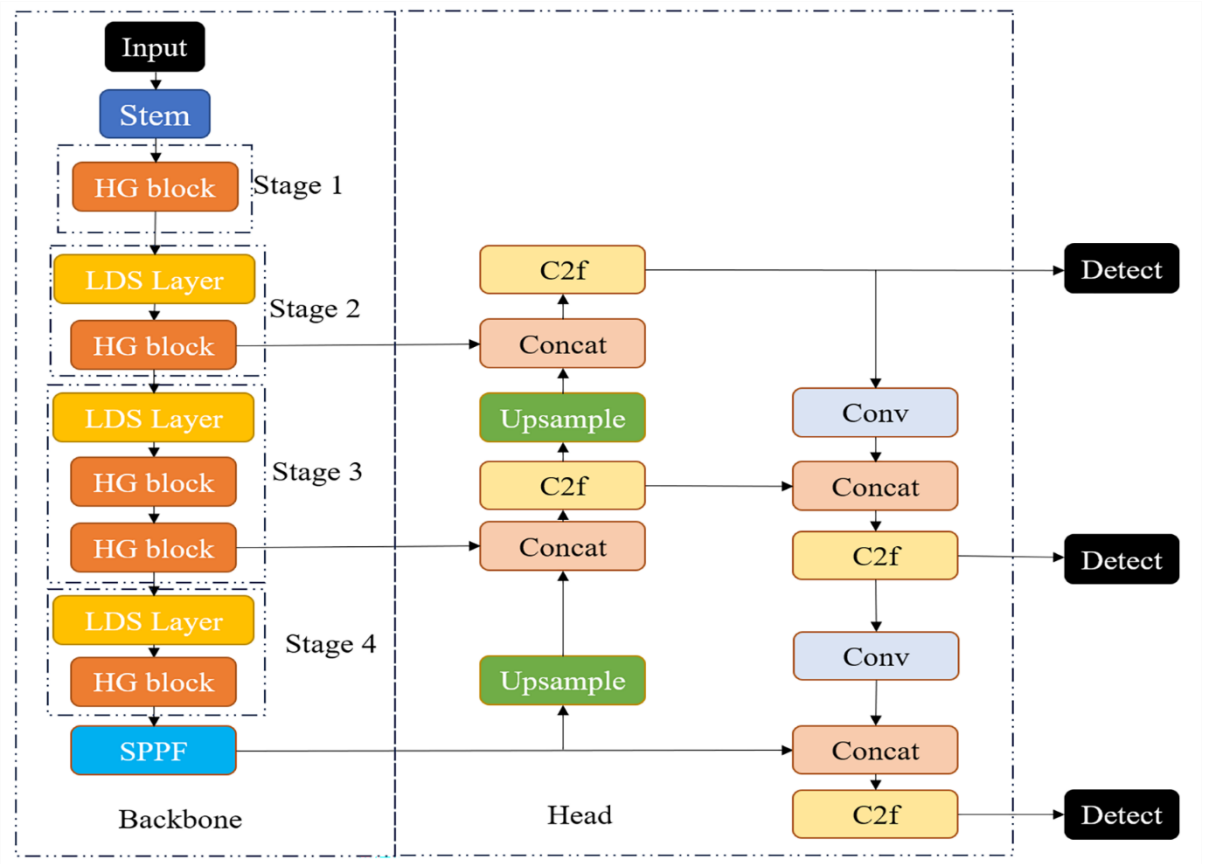

Fig 1. The architecture of HFM-YOLO model

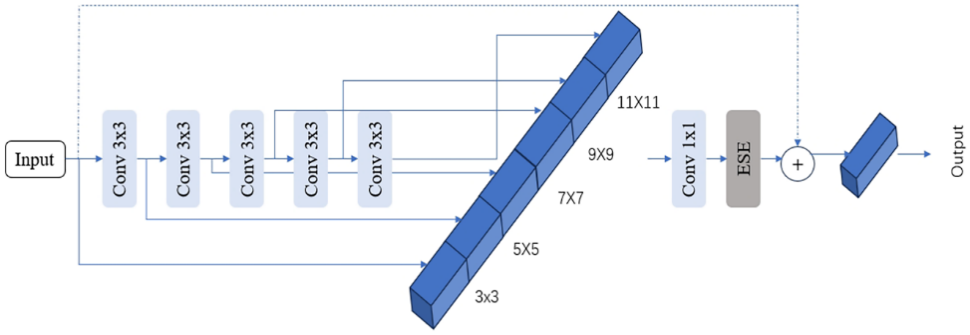

Fig 2. The structure of the HG Block

The core of HGNetV2 is the HG block, which is cleverly designed to process data in a hierarchical manner. This design allows the network to learn from both low-level and high-level features. This results in a rich, multidimensional understanding of the input data. Each HG block is customized to handle different levels of data abstraction. By using HG blocks, the model is able to tell whether the mask is worn or not. This feature is critical for accurate mask detection across different scenarios, including different lighting conditions, angles, and mask types.

The LDS layer is complementary to the HG blocks which is located between the HG blocks. These layers perform the critical downsampling operation, reducing the spatial dimension of the feature map. The LDS layer not only reduces the computational load but also potentially expands the receptive field of subsequent layers. The LDS layer helps improve the overall efficiency of HFM-YOLO, allowing it to process images quickly. The process at high speeds retains the ability to recognize comprehensive feature information. The integration of LDS layer ensures that HFM-YOLO remains computationally efficient, an important property for visual object detection applications. Incorporating these HG blocks and LDS layers into the backbone fundamentally enhances HFM-YOLO's data flow and feature extraction capabilities. The model effectively detects complex layers in visual data and accurately extracts visual features for face mask detection.

HFM-YOLO is further enhanced with the addition of a RepConv layer. The RepConv layer replaces the traditional convolutional layer (Yan, 2023) in the network. These layers are designed to increase efficiency and adaptability, allowing HFM-YOLO to maintain a leaner configuration. At the same time, it can also improve computing output. The RepConv layer excels at dynamically adjusting to different shapes and sizes of masks. This innovation not only reduces the overall computational load of the model but also improves its detection accuracy.

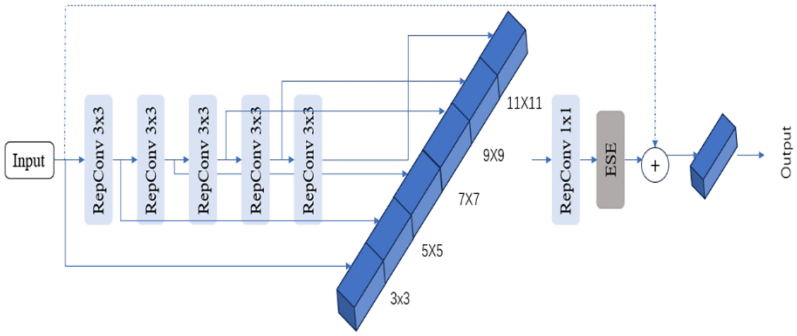

Fig 3. Replacing traditional convolutional layers in the HG block with RepConv

The last part is the Detect part, which is responsible for the final visual object detection.

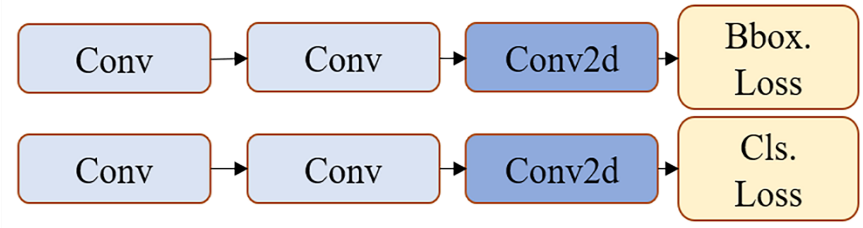

Fig 4. The structure of our detection model

The Bbox Loss in our model is crafted to optimize the accuracy of the predicted bounding boxes in relation to the ground truth. This loss component is bifurcated into two primary segments:

This aspect of the Bbox Loss (Li et al., 2020) concentrates on minimizing the disparity between the central coordinates of the predicted bounding boxes and the ground truth. We employ Mean Squared Error (MSE) (Imran et al.,2019) as a measure for this discrepancy, as it effectively captures the variance in the positional accuracy of the predictions.

$$L_{loss} = \frac{1}{N}\sum_{i=1}^{N}\left(\left(\hat{b}_{xi} - \hat{b}_{xi}\right)^2 + \left(\hat{b}_{yi} - \hat{b}_{yi}\right)^2\right) \quad (1)$$

where, $\hat{b}_{xi}$ and $\hat{b}_{yi}$ are the coordinates of the center, width, and height of the predicted bounding box, and $\hat{b}_{xi}$ and $\hat{b}_{yi}$ are those of the ground truth, with N being the number of bounding boxes.

This segment addresses the differences in the width and height of the predicted bounding boxes compared to the actual dimensions in ground truth (Ju & Cai, 2023). To compute this loss, we utilize the square root of the MSE for the width and height, which aids in balancing the loss contribution from larger versus smaller boxes, ensuring that the model is equally sensitive to the objects with various sizes. Our model employs Class Loss (Cls Loss) (Cui et al., 2019) to fine-tune its capability in accurately classifying the objects. This loss function is crucial for distinguishing between different object categories:

We make use of cross-entropy loss to quantify the deviation between the predicted probability distribution of an object class and the ground-truth distribution. This choice is motivated by using the cross-entropy loss in penalizing inaccuracies in probabilistic predictions. Thus, the classification accuracy of the model is improved.

$$L_{cls} = \frac{1}{N}\sum_{i}^{N}\sum_{c=1}^{C} p_{i,c}\log\left(\hat{p}_{i,c}\right) \quad (2)$$

where C is the number of classes, $p_{i,c}$ is the ground truth probability of class c for the i-th instance, and $p_{i,c}$ is the predicted probability for that class. The incorporation of these specialized loss functions in our model is instrumental in achieving high precision in both object detection and classification tasks. Bbox Loss ensures the accuracy our proposed model in locating objects within an image, while Cls Loss guarantees precise classification.

RESULT

We make use of the face mask detection dataset (Pooja & Preeti, 2021) from Kaggle as the basic data for experimental analysis. The dataset contains 853 images. The dataset was divided into three different categories: Correctly worn masks, not worn masks, and incorrectly worn masks. The data set also covers different scenarios.

Before starting the experiments, we adapted the dataset to a format compatible with the YOLO training format. This preparation involves randomly partitioning the dataset into different subsets for training, validation, and testing purposes. We divided the data set. The training set is taken account for 80% of the total data set, the validation set and test set for 5% and 15% respectively. This distribution is designed to ensure a stable training regime while ensuring complete model evaluation and testing.

Our experiments took use of a Tesla T4 GPU. We used this GPU as the main hardware for the model training, verification and testing stages. The training parameters of all models were standardized to maintain consistency in experimental conditions. Specifically, we configure the training process to 100 epochs and set the batch size to 8.

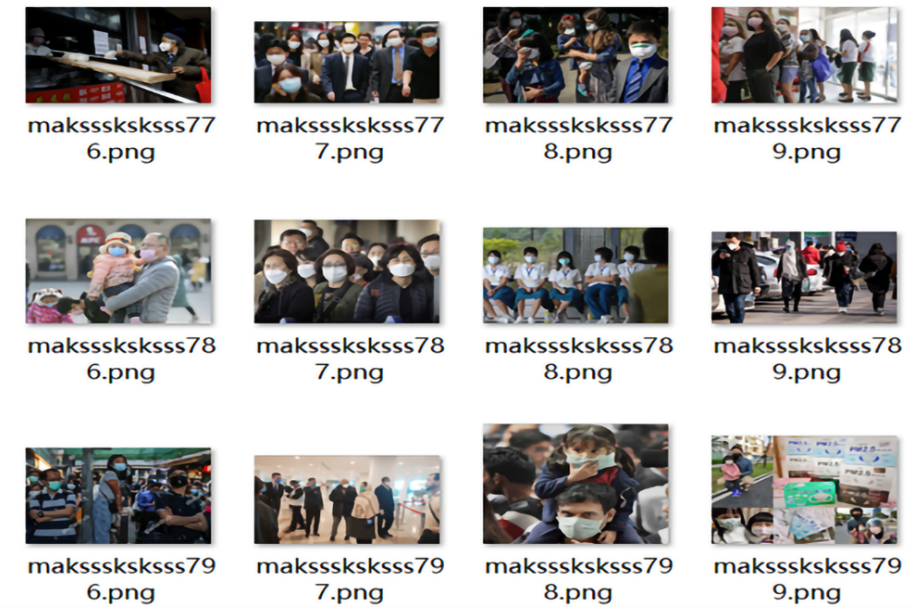

Fig 5. The samples from our dataset

After the training process was completed, we successfully extracted a comprehensive set of experimental results. These results form the basis of our subsequent analysis, providing important insights into the efficacy and performance of the proposed model in the context of mask detection.

In the field of visual object detection, the terms such as true positive (TP), false positive (FP), false negative (FN) and true negative (TN) are widely employed. These metrics are the components of performance evaluation in deep learning because they are directly related to the accuracy and reliability of the model. True positives (TP) are the number of instances that the model correctly identifies as positive. A false positive (FP) refers to a situation where the model incorrectly identifies a negative class instance as a positive class instance. False negatives (FN) represent the number of positive instances that the model failed to identify. True Negatives (TN) A true negative is calculated when the model correctly identifies an object as a negative class.

These metrics are often employed to calculate key performance indicators such as precision (Padilla et al.,2020), recall, and F1 score. Precision measures how accurately a model identifies front-facing objects. In visual object detection, if the bounding box predicted by the model is consistent with the real bounding box, the accuracy is considered correct.

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

Recall rate measures the model's ability to identify all positive objects (Gao et al.,2023). In visual object detection, if the true bounding box coincides with the predicted bounding box, the sample is considered to be correctly recalled.

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

The F1-score is the harmonic average of precision and recall (Liu & Yan, 2021), which provides a single metric to evaluate the overall performance of the model.

$$F1\ Score = 2 \times \frac{Precision \times Reacll}{Precision + Reacll} \tag{5}$$

Through these metrics, we can comprehensively evaluate the performance of object detection models and guide the optimization and application of the models.

Mean Average Precision (mAP) is a metric to determine the performance of an object detection algorithm (Gao et al., 2024). mAP is the average of multiple class average precision (AP), where the AP for each class is calculated from the results of classification. The specific equation is shown as follows,

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{6}$$

where *mAP* is one of the most popular indicators in object detection, it is usually applied to evaluate the accuracy and reliability of object detection algorithms. In order to better evaluate the performance, we make use of mAP50 and mAP50-95 for model evaluation. mAP50 indicates the mAP value if IoU is 0.5. mAP50-95 indicates the mAP value if IoU is 0.5-0.95.

Our Results

The results we obtained are shown in Figure 6.

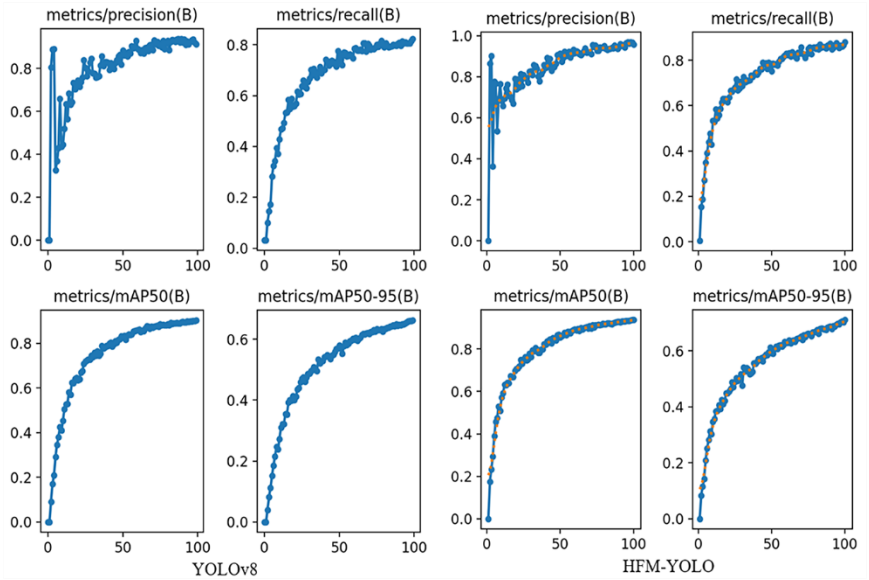

Fig 6. The result of YOLOv8 and HFM-YOLO

From Figure 6, we see that the training losses (Bbox, cls) of all models decrease significantly as epochs increase, indicating that the models improve the predictions over time. The precision of all three models is improved to a very high level, and the recall of the models also increased. This shows that the models were correct in most cases. The accuracy and mAP curves of HFM-YOLO are smoother, which may indicate more stable learning. YOLOv8 has a very high initial validation class loss at the beginning, which may affect performance.

Table 1. Caption text.

Backbone	Precision	Recall	mAP50	mAP50-95
Darknet-19	0.916	0.818	0.903	0.649
CSPDarknet-53	0.913	0.824	0.903	0.662
Darknet-53	0.931	0.865	0.932	0.703
Light-HGNetV2	0.957	0.881	0.938	0.771

In comparison to the conventional Darknet-19 backbone, the implementation of Light-HGNetV2 exhibits a notable improvement of approximately 4.5% in precision. This enhancement is indicative of the model's elevated accuracy in object detection, effectively minimizing the incidence of false positives.

The recall metric, which measures the model's capability to identify all relevant instances, sees Light-HGNetV2 achieving a remarkable rate of 0.881. This number is significantly better than other backbone architectures, highlighting the superior detection capabilities of Light-HGNetV2.

mAP50 is a key metric for evaluating a model's performance at the 50% IoU threshold. Light-HGNetV2 outperforms its peers with a score of 0.938. This is a 3.9% improvement over Darknet-53, the second-highest-performing backbone.

Light-HGNetV2's mAP50-95 demonstrates a substantial leap in performance. It achieves an average precision and recall of 9.7% higher than that achieved by Darknet-53 over the IoU threshold (Yang et al.,2023). This result illustrates the robust and consistent performance of Light-HGNetV2 across varying degrees of object overlap.

The empirical data unequivocally reinforces the effectiveness of the Light-HGNetV2 backbone in the visual object detection framework. The marked improvements in precision, recall, and mean average precision across diverse IoU thresholds attest to the advanced capabilities of Light-HGNetV2 in delivering precise and reliable object detection and classification. These findings compellingly advocate the integration of sophisticated backbone architectures like Light-HGNetV2 within the YOLO framework, paving the way for substantial advancements in the field of visual object detection.

Table 2. Comparison of the model performance in computing speed

Model Names	GFLOPs	FPS
YOLOv5	16.5	52
YOLOv8	8.1	51
Light-HGNetV2	7.7	74

Table 2 clearly shows that our Light-HGNetV2 model, which is the backbone of HFM-YOLO with RepConv integration, requires only 7.7 GFLOPs. This is significantly lower than YOLOv5's 16.5 GFLOPs and even surpasses YOLOv8's 8.1 GFLOPs. This reduction in computational complexity is pivotal, as it indicates a more efficient model that can perform the same tasks with less computational demand.

In terms of FPS, the Light-HGNetV2 backbone enables HFM-YOLO to achieve an impressive 74 frames per second, markedly outperforming YOLOv5's 52 FPS and YOLOv8's 51 FPS. This increase in processing speed is crucial for real-time applications, as it allows for faster detection and response times. This is crucial in scenarios where timely processing is crucial.

RepConv replaces traditional convolutional networks in the HFM-YOLO model. Light-HGNetV2 significantly enhances model performance in terms of computational efficiency and processing speed. Reduce GFLOPs without affecting or even improving FPS. These results demonstrate the use of advanced techniques such as RepConv when developing high-performance, real-time object detection models.

During our testing and evaluation of the HFM-YOLO model, we discovered the model performance that deviated from the expected results. One notable example was a young girl whose mobile phone was identified as not wearing a mask properly as shown in Figure 7.

Fig 7. The detection errors

This particular instance serves as a critical error analysis, highlighting potential directions for improvement of our model. The error highlights challenges to discern subtle differences in mask placement and use. These nuances are critical to public health safety, making accurate detection of mask use imperative.

CONCLUSION

HFM-YOLO represents a major leap forward in object detection, especially in applications involving public health safety, such as mask detection. Our designed Light-HGNetV2 significantly improves detection accuracy and processing efficiency. HFM-YOLO achieved a precision 0.957 in human face mask detection. Although the model demonstrated high efficiency, the challenges pointed out areas that require further enhancement. Overall, HFM-YOLO stands out as a powerful and efficient solution for visual object detection, opening new views for technological advancement in this field.

REFERENCES

Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.

Cui, W., Yan, W. (2016) A scheme for face recognition in complex environments. International Journal of Digital Crime and Forensics (IJDCF) 8 (1), 26-36.

Cui, Y., Jia, M., Lin, T. Y., Song, Y., & Belongie, S. (2019). Class-balanced loss based on effective number of samples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9268-9277).

Gao, X., Nguyen, M., Yan, W. (2021) Face image inpainting based on generative adversarial network. International Conference on Image and Vision Computing New Zealand.

Gao, X., Nguyen, M., Yan, W. (2022) A face image inpainting method based on autoencoder and adversarial generative networks. Pacific-Rim Symposium on Image and Video Technology.

Gao, X., Nguyen, M., Yan, W. (2023) Human face mask detection using YOLOv7+CBAM in deep learning. Handbook of Research on AI and ML for Intelligent Machines and Systems

Gao, X., Nguyen, M., Yan, W. (2023) Enhancement of human face mask detection performance by using ensemble learning models. PSIVT.

Gao, X., Nguyen, M., Yan, W. (2023) A high-accuracy deformable model for human face mask detection. PSIVT

Girshick, R., Donahue, J., Darrell, T., Malik, J., & Mercan, E. (2014). R-CNN for object detection. In *IEEE CVPR*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770-778).

Imran, S., Long, Y., Liu, X., & Morris, D. (2019). Depth coefficients for depth completion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 12438-12447). IEEE.

Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., Kwon, Y., Michael, K., ... & Mammana, L. (2022). ultralytics/YOLOv5: YOLOv5 classification models, apple m1, reproducibility, clearml and deci. ai integrations. *Zenodo*.

Ju, R. Y., & Cai, W. (2023). Fracture detection in pediatric wrist trauma X-ray images using YOLOv8 algorithm. Scientific Reports, 13(1), 20077.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444.

Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., ... & Wei, X. (2022). YOLOv6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*.

Liu, X., & Yan, W. (2021). Traffic-light sign recognition using Capsule network. *Multimedia Tools and Applications*, *80*(10), 15161-15171.

Li, Y., Li, S., Du, H., Chen, L., Zhang, D., & Li, Y. (2020). YOLO-ACN: Focusing on small target and occluded object detection. *IEEE Access*, *8*, 227288-227303.

Liu, M., Yan, W. (2022) Masked face recognition in real-time using MobileNetV2. ACM ICCCV.

Nguyen, M., Yan, W. (2022) Temporal color-coded facial-expression recognition using convolutional neural network. International Summit Smart City 360°: Science and Technologies for Smart Cities.

Nguyen, M., Yan, W. (2023) From faces to traffic lights: A multiscale approach for emotional state representation. IEEE International Conference on Smart City.

Padilla, R., Netto, S. L., & Da Silva, E. A. (2020). A survey on performance metrics for object-detection algorithms. In *International Conference on Systems, Signals and Image Processing (IWSSIP)* (pp. 237-242).

Pollard, C. A., Morran, M. P., & Nestor-Kalinoski, A. L. (2020). The COVID-19 pandemic: A global health crisis. *Physiological Genomics*.

Pooja, S., & Preeti, S. (2021). Face mask detection using AI. Predictive and Preventive Measures for Covid-19 Pandemic, 293-305.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 779-788).

Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, faster, stronger. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7263-7271).

Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Song, C., He, L., Yan, W., Nand, P. (2019) An improved selective facial extraction model for age estimation. International Conference on Image and Vision Computing New Zealand.

Soudy, M., Afify, Y., & Badr, N. (2022). RepConv: A novel architecture for image scene classification on Intel scenes dataset. *International Journal of Intelligent Computing and Information Sciences*, *22*(2), 63-73.

Vedaldi, A., & Zisserman, A. (2016). VGG convolutional neural networks. University of Oxford, 66.

Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7464-7475).

Wang, H., Yan, W. (2022) Face detection and recognition from distance based on deep learning. Aiding Forensic Investigation Through Deep Learning and Machine Learning Framework, pp.144-160, IGI Global.

Xu, G., Yan, W. (2023) Facial emotion recognition using ensemble learning. Deep Learning, Reinforcement Learning, and the Rise of Intelligent Systems, pp.146-158, Chapter 7, IGI Global.

Yan, W. (2019). *Introduction to Intelligent Surveillance: Surveillance Data Capture, Transmission, and Analytics*. Springer.

Yan, W. (2023). *Computational Methods for Deep Learning: Theory, Algorithms, and Implementations*. Springer Nature.

Yang, L., Chen, G., & Ci, W. (2023). Multiclass objects detection algorithm using Darknet-53 and DenseNet for intelligent vehicles. *EURASIP Journal on Advances in Signal Processing*, *2023*(1), 85.

Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., ... & Chen, J. (2024). DETRs beat YOLOs on real-time object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16965-16974).