buildings

MDPI

Article

A Fast and Robust Safety Helmet Network Based on a Mutilscale Swin Transformer

Changcheng Xiang [1,*], Duofen Yin [1], Fei Song [2,*], Zaixue Yu [2], Xu Jian [1] and Huaming Gong [1]

[1] School of Computer Science and Technology, ABA Teachers University , Wenchuan 623000, China; ydf@abtu.edu.cn (D.Y.); jianxu@abtu.edu.cn (X.J.); ghm@abtu.edu.cn (H.G.)
[2] School of Computer and Software Engineering, Xihua University, Chengdu 610209, China; 3120190971450@stu.xhu.edu.cn
* Correspondence: xcc_work@abtu.edu.cn (C.X.); sfei_work@mail.xhu.edu.cn (F.S.)

Abstract: Visual inspection of the workplace and timely reminders of unsafe behaviors (e.g, not wearing a helmet) are particularly significant for avoiding injuries to workers on the construction site. Video surveillance systems generate large amounts of non-structure image data on site for this purpose; however, they require real-time recognition automation solutions based on computer vision. Although various deep-learning-based models have recently provided new ideas for identifying helmets in traffic monitoring, few solutions suitable for industry applications have been discussed due to the complex scenarios of construction sites. In this paper, a fast and robust network based on a mutilscale Swin Transformer is proposed for safety helmet detection (FRSHNet) at construction sites, which contains the following contributions. Firstly, MAE-NAS with the variant of MobileNetV3's MobBlock as a basic block is applied to implement feature extraction. Simultaneously, a multiscale Swin Transformer module is utilized to obtain the spatial and contexture relationships in the multiscale features. Subsequently, in order to meet the scheme requirements of real-time helmet detection, efficient RepGFPN are adopted to integrate refined multiscale features to form a pyramid structure. Extensive experiments were conducted on the publicly available Pictor-v3 and SHWD datasets. The experimental results show that FRSHNet consistently provided a favorable performance, outperforming the existing state-of-the-art models.

Keywords: construction site; MobBlock; mutilscale Swin Transformer; helmet detection

Citation: Xiang, C.; Yin, D.; Song, F.; Yu, Z.; Jian, X.; Gong, H. A Fast and Robust Safety Helmet Network Based on a Mutilscale Swin Transformer. *Buildings* **2024**, *14*, 688. https://doi.org/10.3390/buildings14030688

Academic Editors: Pin-Chao Liao, Xiaowei Luo and Ting-Kwei Wang

Received: 14 January 2024
Revised: 25 February 2024
Accepted: 29 February 2024
Published: 5 March 2024

1. Introduction

Safety helmets are an important labor protection tool in industrial production areas such as construction and manufacturing, and they are widely used and very important [1,2]. However, in real-world scenarios, such as of construction sites or factory assembly lines, many workers still ignore the importance of safety helmets. Simultaneously, because of insufficient corporate supervision, there are countless safety accidents caused by workers entering a site without wearing safety helmets [2]. Therefore, the automatic identification of safety helmets plays a vital role in safe production. By conducting real-time supervision of construction sites, we can sound the alarm for worker safety, while improving worker safety awareness and reducing the occurrence of safety accidents.

In past decades, numerous methods have been presented for the recognition of safety helmet tasks [1,3–5]. Early methods mainly locate safety helmets and workers using sensor and visual attributes (i.e.g, texture, spectral, and structure). Sensor-based helmet detection methods usually adopt Radio Frequency Identification (RFID) [6] tags and readers, e.g., Kelm et al. [4] considered the use of RFID to monitor personnel personal protective equipment [7] compliance. However, the operating range of the RFID reader is an important limiting factor when monitoring safety helmets worn by workers. Visual-attribute-based helmet detection mainly applies machine-learning-based object recognition technologies.

A color-based hybrid descriptor [1] consisting of color histograms [8], hu moment invariants [9], and local binary patterns [10] is presented to claim the feature maps of helmets with different colors (red, yellow, and blue). Subsequently, a hierarchical support vector machine [11] was built to finely classify all of the feature maps into four types of classes (i.e., non-helmet, blue-helmet, yellow-helmet, and red-helmet), thereby achieving a higher accuracy and reliability when processing complex data. Nevertheless, they are heavily dependent on the level of feature design and cannot transfer high-level semantic data.

With the development and progress of computer vision technology in recent years, it has been proven that a series of deep learning methods is very effective at the target recognition of large computer vision datasets—for example, Faster RCNN, the well-known YOLO series [12–14], the Single Shot Multibox Detector (SSD) [15], CornerNet [16], CenterNet [17], and Transformer [18]. However, most approaches are focused on some public data sets with relatively large targets. When identifying an unsafe action while not wearing a safety helmet at a construction site, there may be different visual sizes (small size) of helmets, owing to the different postures of worker and the diversity of safety helmet colors (red, yellow, and blue), see Figure 1, resulting in a low accuracy and high false recognition. For helmet detection, the hybrid deep learning model [19] was applied to combine the convolution neural network (CNN) and the long short-term memory [20]. The SSD-MobileNet algorithm [21] was learned on a dataset containing 3261 images of safety helmets collected from two sources to extract the features of helmets with different colors (red, yellow, and blue). Yu et al. [2] introduced a large-scale, encompassing, and high-quality dataset intended for safety clothing and helmet detection, which adopted some classic object detection methods to verify its effectiveness. However, these classic object detection methods often only focus on the extraction of low-level or high-level features, while ignoring the organic integration between the two, which may affect the reliability and generalization ability of the detection results. Meanwhile, the industry needs to pursue high-performance object detection methods with real-time constraints. To strike the balance between speed and performance, Xu et al. [22] proposed a new detector called DAMO-YOLO, which extends from YOLO but with more new techs, consisting of MAE-NAS backbones, RepGFPN neck, ZeroHead, AlignedOTA and distillation enhancement. Inspired by previous studies, we aim to achieve a fast and accurate helmet detection method via using the respective advantages of Transformer and DAMO-YOLO.

Figure 1. Some examples of safety helmet detection using the SHWD dataset. Note that the helmet comes in different colors.

In this paper, we present a new helmet detection method for promoting the real-time supervision of construction sites. In particular, the contributions are three-fold: (1) We present a fast and robust safety helmet network (FRSHNet) based on mutilscale Swin Transformer for safety helmets to discern the helmet regions of the complex scenarios on the construction site. (2) The Multiscale Swin Transformer (MST) is used to fully extract the available spatial and contextual information in the feature map for each branch of MAE-NAS feature extractor, and efficient-RepGFPN is applied to integrate refined multiscale features to form a pyramid structure. (3) The proposed FRSHNet demonstrates an excellent performance using the publicly available Pictor-v3 and SHWD datasets with the highest mAP of 96.30% and 94.70%, respectively.

The rest of the paper is organized as follows. Section 2 introduces a robust helmet detection network for carrying out real-time supervision of the construction sites. Section 3 validates the recognition performance of the proposed method across two datasets, and finally, Section 4 draws the conclusions.

2. Methods

This section describes a fast and robust safety helmet detection network (FRSHNet) for carrying out real-time supervision of construction sites. Our FRSHNet consists of three main components: (i) Feature extraction based on MAE-NAS; (ii) feature fusion based on a multiscale Swin Transformer and efficient-RepGFPN; and (iii) loss function. The overall procedure of our FRSHNet is shown in Figure 2.

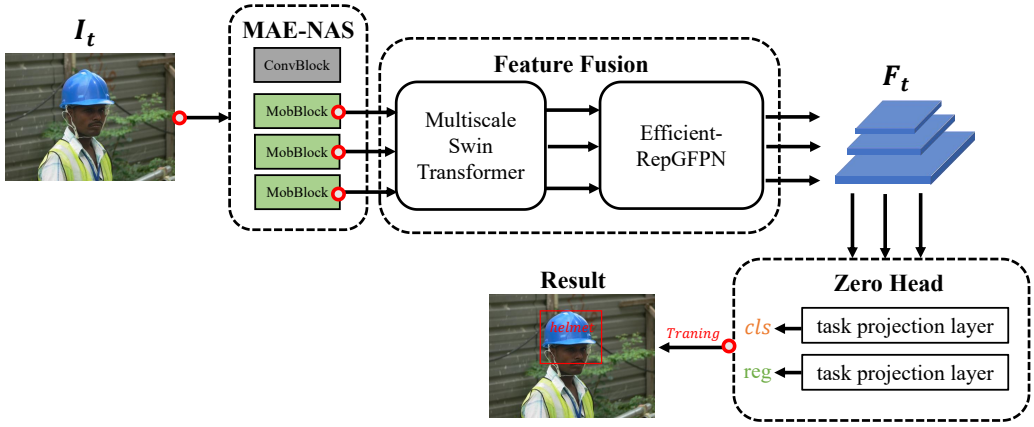

Figure 2. Flowchart of the proposed helmet detection framework. I_t denotes images acquired from the monitoring system at time t, and F_t is a group of multiscale features F_t. MobBlock is a variant of the MobileNetV3 module [23]. The task projection layer only contains a linear layer for regression and a linear layer for classification.

2.1. Feature Extraction Based on MAE-NAS

In a real-world scene at the construction site, we usually need to monitor on-size personal in real time to detect whether they are wearing safety helmets. Previously, in real-time scenarios, researchers depended on the Flops-mAP curve as an easy way to assess the model performance. However, the relationship between the flops and latency of the model was not necessarily consistent. Responding to the principle of latency–MAP curves, DAMO-YOLO proposed that MAE-NAS had the optimal network under different latency budgets.

Let $I_t \in \Re^{H \times W \times 3}$ represent the image obtained from the monitoring system at time t. Because they are affected by the worker's posture, the relative position and shape of the helmet vary greatly. Therefore, we use the MAE-NAS backbones under DAMO-YOLO with different scales to extract basic block. Then, a set of multiscale features F_t of I_t are extracted by the outputs of MobBlocks. As shown in Figure 1, the depths of the three MobBlocks from top to bottom are 96, 128, and 384.

2.2. Feature Fusion Based on the Multiscale Swin Transformer and Efficient-RepGFPN

When the visibility complexity of a construction site is generated by multiple dimensions, including intricate construction equipment, workers' postures and positions, and wearing helmets of different colors and sizes, the performances of helmet detection can be upgraded using the relationships between the worker and helmet [24,25]. In the studies of [26,27], the effectiveness of the Transformer based on the self-attention mechanism when modeling various temporal and spatial position relationships was confirmed. However, when the transformer was applied to vision tasks, its internal spatial self-attention operations led to a significant increase in computational complexity. To this end, the Swin Transformer [18] proposed a novel windowing strategy that grouped the spatial dimensions of the input into multiple non-overlapping windows. In this way, the model only needs to calculate spatial self-attention within each local window, rather than in the entire input space. In response to the above observations, we first integrate the multiscale Swin Transformer module in the feature map of each branch of the MAE-NAS feature extractor, as shown in Figure 3 , to fully leverage spatial positions.

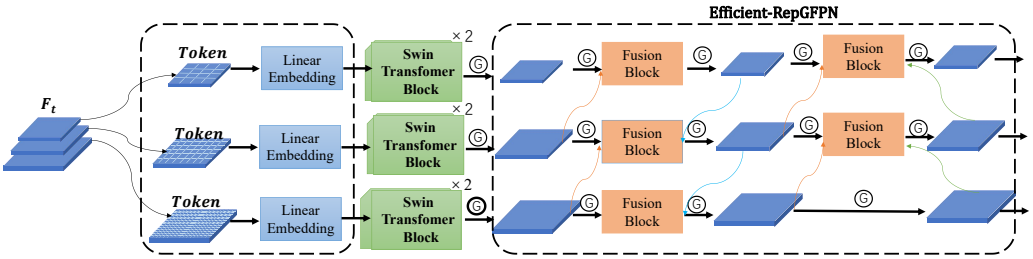

Figure 3. The network architectures of the proposed approach.

Concretely, a patch splitting module was first applied to each feature of the multi-level features F_t to finely extract compact non-overlapping patches. This process ensures that each patch contains meaningful semantic information and is treated as an independent "token". Next, to better process these extracted patches, each patch was transformed through a linear embedding layer. This embedding process enabled each patch to easily be projected to any output dimension, providing greater flexibility for subsequent processing. Subsequently, in order to capture complex patterns and relationships at different scales, three tokens were separately inputt to the Swin Transformer block, which consisted of different layers of (shifted) window-based multi-head self attention ((S)W-MSA) and multilayer perceptron blocks. The advantage of SW-MSA is that it can weigh information at different spatial positions and scales using a weight adjustment mechanism, so the model can adaptively focus on different information sources according to the actual requirements. When calculating MSA, each head is defined as:

$$Att(Q, K, V) = \sigma(\frac{QK^T}{\sqrt{d}} + B)V \quad (1)$$

where $\sigma(\cdot)$ denotes the SoftMax function. Q, K, and V are the core query, key, and value matrices, respectively. The size of these three matrices is $pt^2 \times d$, where d is the channel dimension of the query and key, which determines the number and complexity of features that the model can handle; pt^2 represents the numbers of patches in each window, which is crucial for determining the spatial and context relationship of feature information. $B \in \Re^{pt^2 \times pt^2}$ denotes a relative position bias [18]. In the standard (S)W-MSA, Q, K, and V originate from the same input sequence. The design allows the model to pay different levels of attention to different parts of the input sequence and provide them different weights. In our MSA, Q mainly comes from the corresponding local information f_j^w from each window at a multiscale feature map, and K and V are directly from the token itself. The new design allows the model to simultaneously consider information from features at

different scales and original labels when calculating self-attention. Strictly, for each layer l, we can define our MSA as follows

$$MSA(f^{w}_{j,(l-1)}, T_j) = Concat(head_1, ..., head_h)W^o \quad (2)$$

$$head_k = Att(f^{w}_{j,(l-1)}W^q_j, T_kW^k_j, T_kW^v_j) \quad (3)$$

where W^q_j, W^k_j, $W^v_j \in \Re^{pt^2 \times 512}$, and $W^o \in \Re^{\hat{h}d \times pt^2}$ denote the linear projection matrices that take an input vector and map it to another space using a linear transformation, and $\hat{h}$ is the number of attention heads in a window.

The feature pyramid network aims to perform down-sampling and up-sampling operations on images at different levels, obtain feature maps of different scales, and fuse or cascade these feature maps for subsequent task processing [28]. This has proven to be a critical and effective part of object detection [28,29]. To meet the scheme requirements of real-time target detection—Xu et al. [22] proposed a new efficient RepGFPN that can control the same dimension of the shared channels for each scale feature map under the constraints of limited computational costs, and enhance the feature interaction through queen-fusion. Therefore, after obtaining refined features F_t, we use efficient-RepGFPN to fuse the feature maps of different scales into channels of different sizes as the output features.

2.3. Loss Function

In the training stage, we adopted the distribution focal loss (L_{DFL}) and GIOU (Generalized Intersection over Union) loss (L_{GIOU}) for regression supervision, the quality focal loss (L_{QFL}) for classification supervision. The loss function (L_{helmet}) of FRSHNet is described as

$$L_{helmet} = \alpha L_{DFL} + \beta L_{GIOU} + \eta L_{QFL} \quad (4)$$

where the hyperparameters α, β, and η quantify the contribution of each helmet, respectively. The inference detail of the proposed FRSHNet for a safety helmet detector is summarized in Algorithm 1.

Algorithm 1: A Fast and Robust Safety Helmet Network Based on Mutilscale Swin Transformer

Input: The input images $I_{input} \in \Re^{H \times W \times 3}$.
Output: The bounding box with a score for helmet.
▷ **Feature extraction based on MAE-NAS.**
  Obtain the multiscale features F_t using MAE-NAS with variant MobileNetV3's MobBlock.
▷ **Feature Fusion Based on Multiscale Swin Transformer and Efficient-RepGFPN.**
  Obtain a token by a patch splitting module.
  Compute SW-MSA by Equations (1)–(3).
  Pyramid structure formed by efficient-RepGFPN.
▷ **Loss function.**
  In the training stage, calculate total loss by equation (4).

3. Experimental Results and Discussions

In this section, we consider the proposed FRSHNet and verify its feasibility through experiments. Specifically, we first identify and select a series of representative experimental data sets. Then, implementation details are presented and state-of-the-art models are benchmarked. Finally, we provide a detailed performance analysis of the comparison and ablation experiments.

3.1. Datasets

In our experiments, to effectuate the theoretically feasible goals of FRSHNet, we deliberately selected the Pictor-v3 [30] and SHWD [31] datasets, and carried out validation experiments.

- **Pictor-v3** is a multi-source dataset specifically for helmet detection that fuses images from crowdsourcing (698) and web mining (774). Among these images, the crowd-sourced image contained 2496 worker instances, while the web-mined image contained 2230 worker instances.
- **SHWD** is a public dataset of safety helmet use and head detection, consisting of 7581 high-resolution images. In the SHWD dataset, 9044 safety helmet human subjects were labeled as positive, and 111,514 normal head subjects were labeled as not wearing a helmet or as negative samples.

3.2. Evaluation Criteria

In the helmet detection task, we used the recall, precision, and average precision of each category (i.e, safety helmet or no safety helmet) to evaluate the performance of the proposed method. The recall and precision could comprehensively reflect the detection accuracy and completeness of a helmet detector, respectively. The recall and precision rates are defined as follows:

$$Recall = \frac{N_{tp}}{N_{tp} + N_{fn}} \tag{5}$$

$$Precision = \frac{N_{tp}}{N_{tp} + N_{fp}} \tag{6}$$

where N_{tp} denotes the number of true positives (TPs) that correctly detect the helmet, which measures how accurately the algorithm identifies real helmets. N_{fp} represents the number of false positives (FPs), that is, instances where the algorithm mistakenly believes that a helmet is present. $N_{tp} + N_{fn}$ denotes the total number of all of the ground truth values. It is worth noting that N_{tp} and N_{fp} will vary depending on the setting of the confidence threshold. In order to more comprehensively evaluate the performance of the model, we use a 2D curve (i.e., the precision-recall curve (PRC)) for visualization. On PRC, the abscissa represents the *Recall*, while the ordinate represents *Presision*. The average precision is not simply the average of all precision values, but it is calculated by combining the precision and recall in a specific way. In multi-classification problems, the average accuracy is defined as the area under the PRC, which can more comprehensively reflect the overall performance of the model. The average precision can be formulated as follows:

$$AP = \int_0^1 P(R)dR \tag{7}$$

$$mAP = \frac{1}{N_{cls}} \sum_{i=1}^{N_{cls}} AP_i \tag{8}$$

where AP, P, and R refer to the average precision, precision, and recall, respectively. In our experiments, we mainly applied two metric mAP 0.5 and mAP (0.5:0.95) to evaluate the accuracy and completeness of the proposed method.

3.3. Implementation Details

We implemented the practical aspects of FRSHNet using PyTorch library. The optimizer adopted stochastic gradient descent (SGD), and 64 batch sizes and 100 training epochs were set. The learning rate was initially set to 0.04. In addition, in order to make the experimental results more reliable, the SGD momentum and the weight decay were assigned as 0.9 and $5e^{-4}$ respectively. All of our experiments were carried out on a workstation consisting of a single Nvidia GeForce RTX 4080 GPU and an Intel(R) Core(TM) i9-13900KF CPU.

3.4. Comparison with the State-of-the-Art Models

We estimated the performance of FRSHNet on the Pictor-v3 and SHWD datasets and compared it with seven state-of-the-art target detectors: Faster R-CNN [32], Retina-Net [33], SSD-512 [15], YOLO-v5 [14], FCOS (Fully Convolutional One-Stage Object Detection) [34], VF-Net [35], and TOOD [36]. Table 1 presents the performance of different methods on Pictor-v3 and SHWD datasets. FRSHNet outperformed the other approaches on the Pictor-v3 and SHWD datasets, reaching the highest mAp (Pictor-v3, mAp (0.5) and mAp (0.5:0.95): 96.30% and 65.70%; SHWD, mAp (0.5) and mAp (0.5:0.95): 94.70% and 68.10%), respectively. For the Pictor-V3 dataset and SHWD dataset, TOOD yielded a better performance, as well as a mAp (0.5, Pictor-V3) and a mAp (0.5, SHWD) of 91.50% and 86.70%, respectively. Figures 4 and 5 present some examples of experimental results on the test set.

Figure 4. Visualization of some examples of experimental results generated by the proposed method (FRSHNet) on test set.

Table 1. Comparison results of different object detection methods on the Pictor-v3 and SHWD datasets.

Models	Backbone	Pictor-v3		SHWD	
		mAp (0.50)	mAp (0.5:0.95)	mAp (0.50)	mAp (0.5:0.95)
Faster R-CNN	ResNet-50	0.9060	0.5340	0.8480	0.6310
Retina-Net	ResNet-50	0.9054	0.5438	0.8548	0.6356
SSD-512	VGG16	0.8550	0.4880	0.8080	0.5740
YOLO-v5	CSPDarknet53	0.8818	0.5358	0.8399	0.6386
FCOS	ResNet-50	0.8950	0.5240	0.8580	0.6390
VF-Net	ResNet-50	0.9140	0.5520	0.8570	0.6390
TOOD	ResNet-50	0.9150	0.5580	0.8670	0.6440
FRSHNet	**MAE-NAS**	**0.9630**	**0.6570**	**0.9470**	**0.6810**

Figure 5. Visualization of some examples of the experimental results generated by the proposed method (FRSHNet) on the test set.

3.5. Ablation Studies of FRSHNet

Based on metric learning, FRSHNet includes two main components: (a) feature extraction based on MAE-NAS (MobBlock); (b) feature fusion based on multiscale Swin Transformer and efficient RepGFPN. Therefore, we designed ablation experiments on FRSHNet to test and verify the validity of (a) and (b) modules.

- FRSHNet$^{\#1}$: ResNet18 + (b)+ Zero Head.
- FRSHNet$^{\#2}$: ResNet50 + (b)+ Zero Head.
- FRSHNet$^{\#3}$: (a) + Zero Head.
- FRSHNet: (a) + (b) + Zero Head.

Table 2 shows the ablation studies of the baseline and its different variants on the SHWD test set. From the table, we can clearly observe that when modules (a) and (b) were enabled at the same time, the performance of FRSHNet was significantly improved compared with the baseline. This means that the merger of these two modules resulted in more comprehensive and coordinated performance enhancements in the model. In addition, we also noted certain improvements over the baseline when (a) or (b) modules were enabled individually, which further confirmed the effectiveness of each module.

Table 2. Ablation studies of FRSHNet on SHWD datasets. FPS (Frames Per Second) represents the average number of images that can be processed by the model per second.

Models	mAp (0.50)	mAp (0.5:0.95)	FPS
FRSHNet$^{\#1}$	0.9170	0.5750	6873.3
FRSHNet$^{\#2}$	0.9320	0.6210	6634.6
FRSHNet$^{\#3}$	0.8520	0.5270	7342.1
FRSHNet	**0.9470**	**0.6810**	6785.5

Specifically, when compared with FRSHNet$^{\#1}$, FRSHNet$^{\#2}$ enhanced mAP (0.5) and mAP (0.5:0.95) by 1.5 % and 4.6 % on the SHWD dataset, respectively. Furthermore, mAP (0.5) and mAP (0.5:0.95) were obtained as 85.20 % and 52.70% on the SHWD dataset when integrating (a) and Zero Head modules (i.e., FRSHNet) to baseline, i.e., FRSHNet$^{\#3}$. Compared with FRSHNet$^{\#3}$, mAP (0.5) and mAP (0.5:0.95) of FRSHNet integrating (a) and (b) modules were upgraded by 9.5% and 15.4%. The great improvement of the FRSHNet further demonstrates the effectiveness of (a) and the modules. In addition, as these models integrate the distributed reasoning principle of DAMO-YOLO, their FPS (Frames Per Second) was very high. Figure 6 presents the PRC for FRSHNet on the SHWD test set.

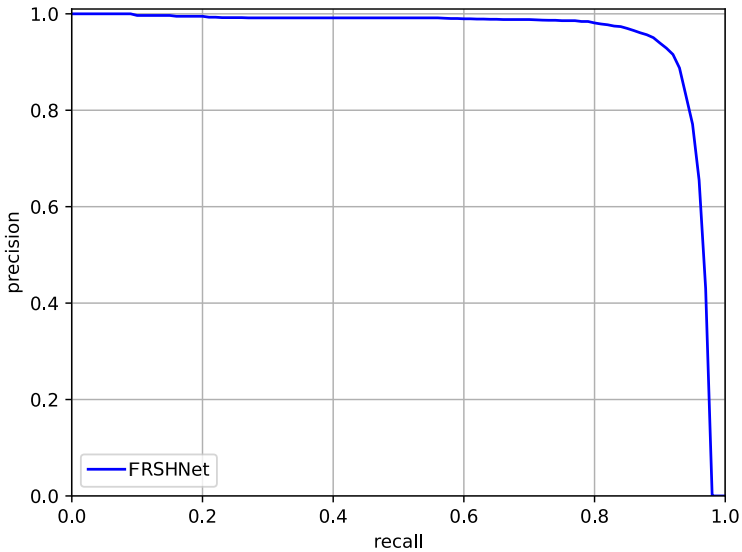

Figure 6. Precision–Recall curve (PRC) for FRSHNet on the SHWD test set.

4. Conclusions

Visual inspection of the workplace and timely reminders of unsafe behaviors (e.g., not wearing a helmet) regarding safety helmets are very important safety management measures to avoid injuries to workers on construction sites. Video surveillance systems generate large amounts of non-structure image data on site for this purpose, but they require real-time recognition automation solutions based on computer vision. Although various deep-learning-based models have recently provided new ideas for helmet recognition in traffic monitoring, few solutions suitable for industry applications have been discussed owing to the complexity of the construction site scenarios. This paper described a fast and robust network FRSHNet based on a mutilscale Swin Transformer that was proposed to identify safety helmets at the construction site. Three key contributions of our method adopted MAE-NAS with the variant MobileNetV3's MobBlock for feature extraction, multiscale Swin Transformer Module for generating the spatial and context relationships, and efficient-RepGFPN for real-time detection. The results from the Pictor-v3 and SHWD datasets showed that the proposed method had advantages over other state-of-the-art methods.

In future work, we will carry out the following related work: (i) dataset diversity and real-world testing, including images from monitored construction sites with varying lighting and weather conditions, to enhance the robustness and practical applicability of the proposed method; (ii) model optimization for energy efficiency and reduced computational load, considering that it is crucial for real-time applications and possible deployment on portable devices with limited processing capabilities; and (iii) expand the model's capabilities to detect other types of personal protective equipment, such as safety vests and goggles, for comprehensive workplace safety monitoring.

Author Contributions: C.X.: conceptualization, formal analysis, methodology, model building, writing—original draft, writing—review and editing. D.Y.: conceptualization, formal analysis, model building writing—review and editing. F.S.: conceptualization, model building, validation, writing—review and editing. Z.Y.: model building, and validation. X.J.: validation, writing—original draft, writing—review and editing. H.G.: formal analysis, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Informed consent was obtained from all of the subjects involved in the study.

Data Availability Statement: Pictor-v3 [30] and SHWD [31] are two publicly available helmet recogintion datasets at: https://github.com/ciber-lab/pictor-ppe (accessed on 4 January 2021) and https://github.com/njvisionpower/Safety-Helmet-Wearing-Dataset (accessed on 1 December 2019).

Acknowledgments: The authors would like to thank all of the researchers who kindly sharing the codes and datasets used in our studies.

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

Abbreviations

The following abbreviations are used in this manuscript:

Abbreviations	Description
RFID	Radio frequency identification
CNN	Convolution neural network
FOCS	Fully convolutional one-stage object detection
SSD	Single shot multibox detector
(S)W-MSA	(Shifted)-window-based multi-head self attention
ViT	Vision transformer
PRC	Precision–recall curve

References

1. Wu, H.; Zhao, J. An intelligent vision-based approach for helmet identification for work safety. *Comput. Ind.* **2018**, *100*, 267–277. [CrossRef]
2. Yu, F.; Wang, X.; Li, J.; Wu, S.; Zhang, J.; Zeng, Z. Towards Complex Real-World Safety Factory Inspection: A High-Quality Dataset for Safety Clothing and Helmet Detection. *arXiv* **2023**, arXiv:2306.02098 .
3. Chen, C.; Wu, W. Color pattern recognition with the multi-channel non-zero-order joint transform correlator based on the HSV color space. *Opt. Commun.* **2005**, *244*, 51–59. [CrossRef]
4. Kelm, A.; Laußat, L.; Meins-Becker, A.; Platz, D.; Khazaee, M.J.; Costin, A.M.; Helmus, M.; Teizer, J. Mobile passive Radio Frequency Identification (RFID) portal for automated and rapid control of Personal Protective Equipment (PPE) on construction sites. *Autom. Constr.* **2013**, *36*, 38–52. [CrossRef]
5. Li, Y.; Wei, H.; Han, Z.; Huang, J.; Wang, W. Deep Learning-Based Safety Helmet Detection in Engineering Management Based on Convolutional Neural Network. *Adv. Civ. Eng.* **2020**, *2020*, 10. [CrossRef]
6. Rajaraman, V. Radio frequency identification. *Reson* **2017**, *22*, 549–575. [CrossRef]
7. Dolez, P.I. Chapter 3.6—Progress in Personal Protective Equipment for Nanomaterials. In *Nanoengineering*; Dolez, P.I., Ed.; Elsevier: Amsterdam, The Netherlands, 2015; pp. 607–635. [CrossRef]
8. Swain, M.J.; Ballard, D.H. Indexing via Color Histograms. In Proceedings of the Active Perception and Robot Vision, Maratea, Italy, 16–29 July 1989; Sood, A.K., Wechsler, H., Eds.; Springer: Berlin/Heidelberg, Germany, 1992; pp. 261–273.
9. Žunić, J.; Hirota, K.; Rosin, P.L. A Hu moment invariant as a shape circularity measure. *Pattern Recognit.* **2010**, *43*, 47–57. [CrossRef]
10. Pietikäinen, M. Local Binary Patterns. *Scholarpedia* **2010**, *5*, 9775. [CrossRef]
11. Zhigang, L.; Wenzhong, S.; Qianqing, Q.; Xiaowen, L.; Donghui, X. Hierarchical support vector machines. In Proceedings of the 2005 IEEE International Geoscience and Remote Sensing Symposium (IGARSS '05), Seoul, Republic of Korea, 20 July 2005; Volume 1, p. 4. [CrossRef]
12. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
13. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
14. Jocher, G.; Stoken, A.; Borovec, J.; NanoCode012, C.; Changyu, L.; Laughing, H. ultralytics/yolov5: v3.0. 2020. Available online : https://github.com/ultralytics/yolov5 (accessed on 20 December 2020).
15. Liu, W .; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
16. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750.
17. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850 .
18. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv* **2021**, arXiv:2103.14030 .
19. Ding, L.; Fang, W.; Luo, H.; Love, P.E.; Zhong, B.; Ouyang, X. A deep hybrid learning model to detect unsafe behavior: Integrating convolution neural networks and long short-term memory. *Autom. Constr.* **2018**, *86*, 118–124. [CrossRef]
20. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
21. Othman, N.A.; Aydin, I. A New Deep Learning Application Based on Movidius NCS for Embedded Object Detection and Recognition. In Proceedings of the 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkey, 19–21 October 2018; pp. 1–5. [CrossRef]
22. Xu, X.; Jiang, Y.; Chen, W.; Huang, Y.; Zhang, Y.; Sun, X. DAMO-YOLO: A Report on Real-Time Object Detection Design. *arXiv* **2023**, arXiv:cs.CV/2211.15444.
23. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
24. Chen, H.; Shi, Z. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote. Sens.* **2020**, *12*, 1662. [CrossRef]
25. Song, F.; Zhang, S.; Lei, T.; Song, Y.; Peng, Z. MSTDSNet-CD: Multiscale Swin Transformer and Deeply Supervised Network for Change Detection of the Fast-Growing Urban Regions. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
26. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the Computer Vision—ECCV 2020 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 213–229.
27. Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; Gao, W. Pre-trained image processing transformer. *arXiv* **2020**, arXiv:2012.00364.
28. Ghiasi, G.; Lin, T.Y.; Le, Q.V. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7036–7045.

29. Jiang, Y.; Tan, Z.; Wang, J.; Sun, X.; Lin, M.; Li, H. GiraffeDet: A heavy-neck paradigm for object detection. *arXiv* **2022**, arXiv:2202.04256 .
30. Nath, N.D.; Behzadan, A.H.; Paal, S.G. Deep Learning for Site Safety: Real-Time Detection of Personal Protective Equipment. *Autom. Constr.* **2020**, *112*, 103085. [CrossRef]
31. Gochoo, M. Safety Helmet Wearing Dataset. Mendeley Data. 2021. Available online : https://github.com/njvisionpower/Safety-Helmet-Wearing-Dataset (accessed on 17 December 2019).
32. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
33. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2999–3007. [CrossRef]
34. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9627–9636.
35. Zhang, H.; Wang, Y.; Dayoub, F.; Sünderhauf, N. VarifocalNet: An IoU-aware Dense Object Detector. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 8510–8519. [CrossRef]
36. Feng, C.; Zhong, Y.; Gao, Y.; Scott, M.R.; Huang, W. TOOD: Task-aligned One-stage Object Detection. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 3490–3499. [CrossRef]