remote sensing

MDPI

Article

A Remote-Vision-Based Safety Helmet and Harness Monitoring System Based on Attribute Knowledge Modeling

Xiao Wu, Yupeng Li, Jihui Long, Shun Zhang *, Shuai Wan and Shaohui Mei

School of Electronic and Information, Northwestern Polytechnical University, Xi'an 710129, China
* Correspondence: szhang@nwpu.edu.cn

Abstract: Remote-vision-based image processing plays a vital role in the safety helmet and harness monitoring of construction sites, in which computer-vision-based automatic safety helmet and harness monitoring systems have attracted significant attention for practical applications. However, many problems have not been well solved in existing computer-vision-based systems, such as the shortage of safety helmet and harness monitoring datasets and the low accuracy of the detection algorithms. To address these issues, an attribute-knowledge-modeling-based safety helmet and harness monitoring system is constructed in this paper, which elegantly transforms safety state recognition into images' semantic attribute recognition. Specifically, a novel transformer-based end-to-end network with a self-attention mechanism is proposed to improve attribute recognition performance by making full use of the correlations between image features and semantic attributes, based on which a security recognition system is constructed by integrating detection, tracking, and attribute recognition. Experimental results for safety helmet and harness detection demonstrate that the accuracy and robustness of the proposed transformer-based attribute recognition algorithm obviously outperforms the state-of-the-art algorithms, and the presented system is robust to challenges such as pose variation, occlusion, and a cluttered background.

Keywords: automated safety checking system; safety helmets and harnesses; attribute recognition based on transformer; construction site datasets

check for updates

Citation: Wu, X.; Li, Y.; Long, J.; Zhang, S.; Wan, S.; Mei, S. A Remote-Vision-Based Safety Helmet and Harness Monitoring System Based on Attribute Knowledge Modeling. *Remote Sens.* **2023**, *15*, 347. https://doi.org/10.3390/rs15020347

Academic Editor: Pablo Rodríguez-Gonzálezvez

Received: 8 November 2022
Revised: 19 December 2022;
Accepted: 27 December 2022
Published: 6 January 2023

1. Introduction

The five major types and causes of accidents occurring on construction sites are: falling from a height; being struck by objects; mechanical and hoisting damage; electrocution; and collapse. The death toll of these five construction fatalities accounts for over 90% of all fatal incidents in the construction industry. The fatal incidence of falling from a height is the highest among these causes, and the safety risk is also exorbitant [1]. The U.S. Occupational Safety and Health Administration (OSHA) and similar agencies in other countries aim to develop and impose rules and regulations on construction sites to reduce injuries. They found that all personnel working in close proximity to site hazards should wear appropriate personal protective equipment (PPE) to minimize the risk of being exposed to or injured by hazards [2]. For example, a safety helmet and safety harness, which are the most common PPE components, can absorb and diffuse the impact of falling, reducing the risk of injury to workers who fall from heights. However, for various reasons, such as workers' simple negligence or misinformation, these two PPE components are not always worn properly. Hence, as a preventive step, an automatic safety helmet and harness monitoring system is critical for construction contractors to enforce worker-safety monitoring.

With the development of computer vision technology, almost all automatic monitoring methods for safety helmets and harnesses based on video streams are object detection problems and are solved using computer-vision-based techniques. Among them, deep-learning-based methods with convolutional neural networks (CNN) have made significant breakthrough progress in object detection owing to the advantage of extracting deep and high-level feature representations from raw image pixels [3] and reducing the effort in modeling prior knowledge of interest manually. Many studies have applied CNN-based methods [4–7] for safety helmet and harness detection, which locate the target object and identify a particular target's category. For example, Han et al. [4] presented an object detection algorithm based on a single-shot multibox detector (SSD) to solve the problem of low accuracy in existing safety helmet detection. In [5], a hierarchical positive sample selection (HPSS) mechanism was proposed to improve the fitting ability of YOLOv5 for efficient safety helmet detection. The work in [8] adopted the object detection network YOLOv5 and the human body posture estimation network OpenPose for the detection of safety harnesses. A computer-vision-based approach for safety harness detection, ref. [9] used a Faster-R-CNN to detect the presence of a worker and a deep CNN model to determine if workers were wearing their harnesses when performing tasks while working at heights. Despite the great success of deep-learning-based techniques for safety helmets and harness detection, they have two limitations. On the one hand, object detection is not ideal for the recognition of small targets and occlusions, which is demonstrated in detail later. On the other hand, object detection methods are often trained using large-scale datasets in a fully supervised manner, whereas there are relatively few public datasets available for a detailed evaluation of safety helmet and harness monitoring systems. When applied in actual scenarios, we have to consider how to alleviate these problems.

Motivated by the operating procedures of human experts and recent research in attribute learning [10], semantic attribute representations (such as gender, hairstyle, or clothing style) are reliable and robust to the variance of workers' appearance in unknown poses. In this study, we introduce a new class of midlevel attributes related to construction safety states and transfer the traditional detection problem of safety helmets and harnesses into a semantic attribute recognition problem of construction safety states. We designed a novel attribute knowledge modeling network based on the transformer architecture, in which the self-attention mechanism is applied to fully explore the relationship between semantic attributes and image features for attribute recognition. Using the algorithms of detection, tracking, and our proposed attribute knowledge modeling, a safety recognition system and a real-time human–computer interface for use in construction sites are presented. The system can intelligently identify whether workers comply with the safety regulations and specifications. We collected the video streams of workers wearing safety helmets and harnesses to create an open-site monitoring dataset for construction scenes, which contains three subdatasets: object detection, multiobject tracking, and attribute recognition. The experimental results prove that the mean accuracy (mA) of our attribute recognition model in recognizing safety helmets and harnesses is 96%, which has a high application value.

We make the following contributions to this work:

- We propose an automatic safety helmet and harness monitoring system based on attribute knowledge modeling to recognize the wearing states of safety helmets and harnesses. In contrast to previous studies that apply object detection to locate and identify safety helmets and harnesses, we transfer this problem into a semantic attribute recognition problem, which is more reliable and robust to the variance in workers' appearances in unknown poses.
- We present a novel attribute knowledge modeling network based on the transformer architecture, in which the self-attention mechanism is applied to fully explore the relationship between attributes and image features for attribute recognition.
- We develop an open-site monitoring dataset for construction scenes containing three subdatasets: object detection, multiobject tracking, and attribute recognition. This benchmark dataset is crucial for evaluating safety helmet and harness monitoring systems in unconstrained environments.

The remainder of this study is organized as follows. Section 2 introduces research related to this study. In Section 3, we describe the proposed safety helmet and harness monitoring system in detail. Section 4 introduces the utilized datasets and demonstrates the implementation details of our experiments. The experimental results are presented in

Section 5. A discussion of our work is in Section 6. Finally, in Section 7, conclusions and future work are summarized.

2. Related Works

In this section, we provide background knowledge and review related work on safety helmet and harness monitoring systems, pedestrian attribute recognition, and transformers in computer vision.

2.1. The Safety Helmet and Harness Monitoring System

The safety helmet and harness monitoring system at construction sites plays a crucial role for electronic eyes in safeguarding workers. Recent safety helmet and harness monitoring systems can be divided into two classes: computer-vision-based [11,12] and wearable-sensor-based methods [13,14]. Although numerous wearable-sensor-based methods use contact sensors to gather data effectively, these sensors are often expensive, precluding their widespread use. In contrast, computer-vision-based techniques are advantageous as noncontact optical techniques that can be robust, hygienic, reliable, safe, cost-effective, and suitable for long-distance and long-term monitoring. In terms of computer-vision-based methods, Zdenek et al. [15] investigated how to improve the safety at construction sites using CNN models for safety guardrail detection. This work was inspired by the fact that most construction accidents are caused by falls from heights due to unguarded edges. Fang et al. [16] proposed a method for automatically detecting the personal protective equipment of construction workers. It adopted the Faster R-CNN algorithm to detect bareheaded workers from field images with high speed and accuracy. Nath et al. [6] applied CNNs to detect multiple pieces of personal protective equipment, such as hard hats and safety vests, from surveillance videos. Similarly, SSD and CNN were suggested by Wu et al. [7] for detecting construction personnel wearing hard hats. Recently, Shanti et al. [17] developed a novel technique that monitored whether the workers were complying with the safety standard of the Personal Fall Arrest System (PFAS). The real-time detection algorithms they built included safety helmets, safety harnesses, and lifeline. With the development of UAV emergency monitoring [18], Shanti et al. [19] also focused on UAVs, and proposed the use of UAVs to monitor workers in real-time while performing high-altitude activities.

With the recent development of deep-learning-based object detection approaches, the safety helmet and harness detection has achieved breakthrough performance. However, the problem that has to be considered in the application is that the target cannot be detected due to its small scale and occlusion in the actual scene. In our system, we innovatively propose to transform the detection problem into the problem of semantic attribute recognition of images to alleviate this deficiency. Considering the economy, we did not choose the superior performance of the drone. We choose to use the existing monitoring system on the construction site and only need to deploy a high-performance GPU server to realize the remote-vision-based safety helmet and harness monitoring system.

2.2. Pedestrian Attribute Recognition

Earlier pedestrian attribute recognition methods generally modeled each attribute independently based on hand-crafted features, such as color histograms and texture histograms [20,21]. With the success of deep learning, pedestrian attribute recognition has gained considerable attention in recent years, and many pedestrian attribute recognition approaches based on deep networks [22–24] have been developed. Most of these methods utilize a CNN or attention mechanisms to capture discriminative features [23,25]. Li et al. [26] treated the pedestrian attribute recognition task as a multilabel classification task [27,28] and designed a weighted sigmoid cross-entropy loss to relieve the unbalanced attribute problem. HydraPlus-Net [25] was introduced to encode multiscale features from multiple levels for pedestrian analysis using the multidirectional attention (MDA) mechanism.

Recently, some researchers have gradually focused on exploring the relationship between image regions and semantic attributes. Sarfraz et al. [29] constructed their models based on multitask learning (MTL), which learned the commonalities of all attributes, but ignored the individuality of each attribute. A CNN–RNN-based encoder–decoder framework was proposed in JRL [30], which aimed to discover the interdependence and correlation among attributes with an LSTM model. GRL [31] split the body into regions and fed the features of regions into the RNN to explore the correlations of the regions. Zhao et al. [22] proposed two models, i.e., recurrent convolutional (RC) and recurrent attention (RA) models. They explored the correlations between different attribute groups, including the intragroup spatial locality correlation and the intergroup attention correlation through a convolutional-LSTM network, respectively. Li et al. [32] performed reasoning using graph convolutional networks (GCNs), in which one graph captured spatial relations between regions and the other one learned potential semantic relations between attributes. In JLAC [33], Tan et al. applied a GCN to build an attribute graph of attribute-specific features and explored contextual relations.

As most previous methods consider the independence of attributes, they achieve poor performance on pedestrian attribute recognition owing to their failure to exploit relations between regions and attributes. In contrast, our work aims to extract discriminative features from the transformer architecture and consider capturing the relationship between the semantic attributes and spatial features.

2.3. Transformer in Computer Vision

Transformer [34] was first proposed to model long-range dependencies in sequence-learning problems and has been widely used in natural language processing (NLP) tasks [35–38]. Recently, transformer-based models have been applied to many computer vision (CV) tasks [39–41] and have shown great potential. Dosovitskiy et al. [39] proposed vision transformers (ViT), in which they split an image into multiple patches and fed them into a stacked transformer architecture for classification. Carion et al. [40] designed an end-to-end object detection framework named DETR with a transformer, and it achieved a good performance in object detection. For the task of object reidentification, He et al. [41] proposed TransReID. The side-information embedding was plugged to encode camera or viewpoint information, and a jigsaw patches module was designed to learn more robust features. The success of transformers can be mainly attributed to self-supervision and self-attention [42]. The self-supervision can train complex models without the high cost of human annotation and encode useful relationships between entities presented. The self-attention takes the context of a given sequence into account by learning the relations between the elements in the token set (e.g., words in language or patches in an image). Some methods [43–45] demonstrate the potential of the transformer architecture in capturing sequence relationships.

Our study is inspired by the DETR method [40] for object detection. However, unlike most existing works, we introduce a transformer to address the problem of attribute recognition and leverage the transformer to extract spatial and semantic information between features and attributes. Moreover, we exploit the self-attention mechanism to learn attribute relations to improve the feature representations for a higher accuracy performance.

3. Methodology

3.1. Overall System

To achieve intelligent security monitoring, we designed an automatic safety helmet and harness monitoring system for construction sites. The proposed system is based on worker detection, tracking, and safety working state identification based on attribute knowledge modeling. The overall framework of the system is shown in Figure 1.

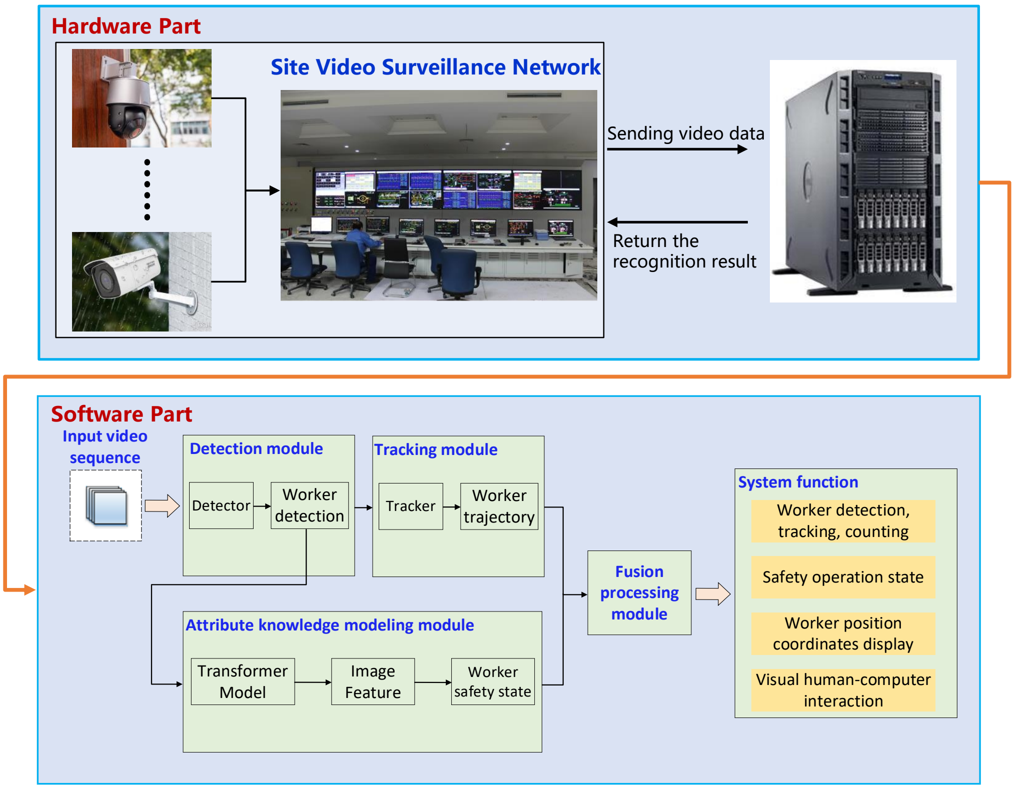


Figure 1. The framework of our proposed system. The system is divided into the hardware part and the software part. The hardware part is used to send input video data and display the recognition results; the software part is used to intelligently process video and images.

The overall system was divided into two parts: hardware and software. The hardware part included a video surveillance network on construction sites and a GPU high-performance image processor. The video surveillance network at construction sites contains a camera network and a monitoring computer. The video stream data were first collected by the camera network and then sent by the monitoring computers to the GPU high-performance image processor for vision-based image processing. After processing, the recognition and warning results were returned to the monitoring computer. For the software part, the input image sequence was input into the detection module for worker detection and the tracking module for worker tracking. To identify a worker's safety state, the attribute knowledge modeling module treated the safety states as semantic attributes and applied a transformer for attribute recognition. Finally, we introduced a fusion processing module to integrate the tracking results and attribute recognition results. Next, we introduce the specific modules in detail.

3.2. Worker Detection and Tracking

The algorithm for worker detection and tracking is shown in Figure 2. We selected the common YOLOv5 as the detector and Deep SORT as the tracker. First, the video was input into the YOLOv5 detector to locate workers in each video frame, and then we extracted the bounding box and feature map for each detection. For Deep SORT, we extracted the motion and surface features with two branches. For the motion feature extraction, we calculated the Markov distance according to the bounding boxes of workers in continuous frames, extracted the motion features of the particular target through the Kalman filter, and applied the Hungarian algorithm to match two adjacent frames. For the surface feature extraction, a CNN model was used to extract appearance feature information. Finally, the final tracking result was acquired by combining the motion and surface features.

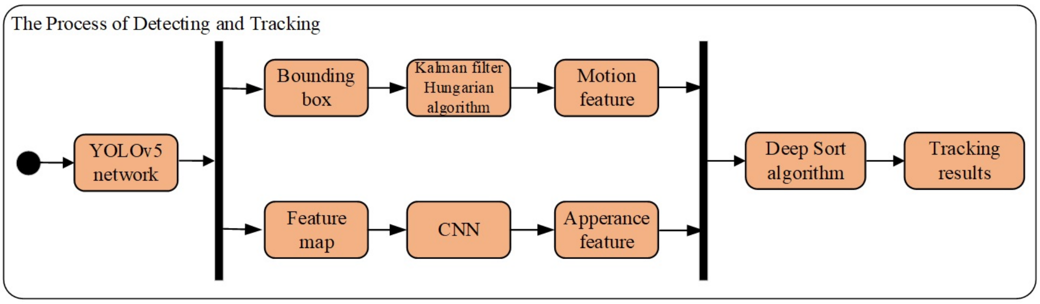


Figure 2. The process of detecting and tracking. We adopted YOLOv5 as the detector and extract motion features and appearance features based on the Kalman filter and Hungarian algorithm to input into the Deep SORT algorithm to get tracking results.

3.3. Safety Attribute Recognition

In this study, the problem of recognizing a safe working state (such as wearing safety helmets and harnesses) on construction sites was transferred to the problem of the recognition of images' semantic attributes. Here, we present the utilization of the self-attention mechanism of the transformer for the safety-attribute recognition, as shown in Figure 3. It can be seen our method can be divided into four parts, namely, feature and attribute embeddings, relation exploitation based on a transformer, a classifier, and a loss function. We treated each feature embedding or attribute label embedding as a word vector and input the transformer encoder for mutual learning simultaneously. Then, utilizing the transformer encoder, we computed the dependencies between attributes and features by self-attention. We finally obtained a set of weights rich in spatial and semantic information and the final prediction results and losses.

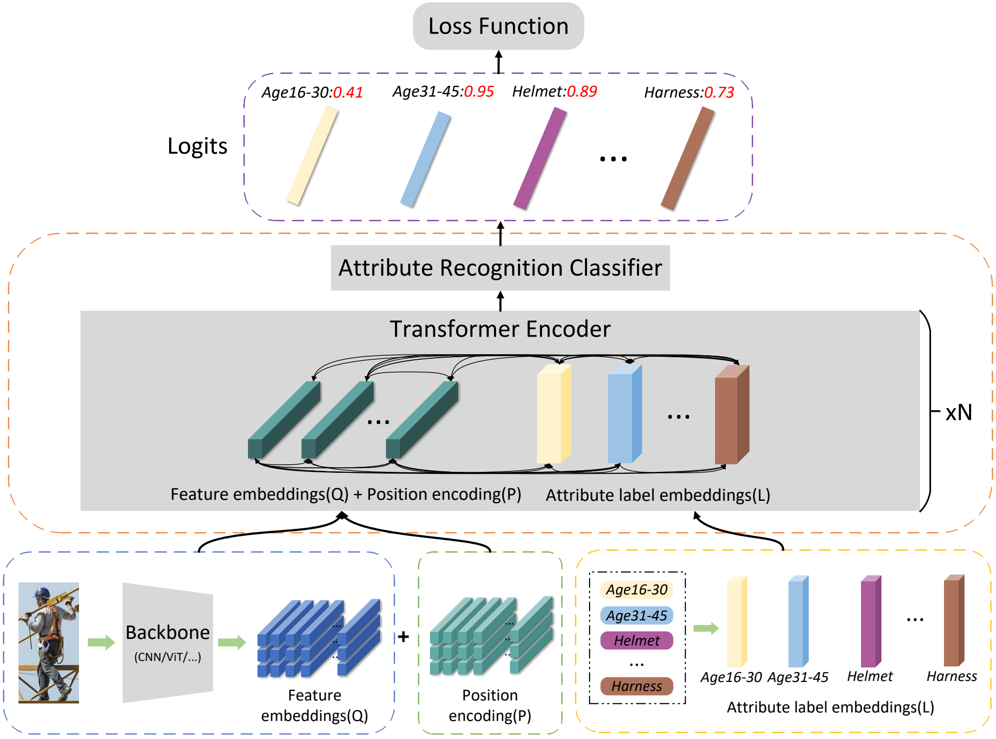


Figure 3. The framework of our proposed method. We use the transformer encoder to model the relationship between feature embeddings and attribute label embeddings.

3.3.1. Feature and Attribute Embeddings

Image Feature Embeddings Q**.** Given an input image $x \in \mathbb{R}^{H \times W \times 3}$, the feature extractor (e.g., ResNet in Figure 3) output a tensor $Q \in \mathbb{R}^{h \times w \times d}$, where h, w, and d were the output height, width, and channel, respectively. We took the tensor Q as a set of vectors, where $Q = \{q_1,q_2,\ldots,q_n\}$, $q_i \in \mathbb{R}^d$, with i ranging from 1 to n (where $n = h \times w$). Thus, we had a tensor $Q \in \mathbb{R}^{n\times d}$, in which each feature embedding represented a subregion that mapped back to a patch in the original image space. At the same time, we ran it through an embedding layer to initialize the extracted image's local feature tensor Q to generate a learnable position encoding $P = \{p_1,p_2,\ldots,p_n\}$, where

$$p_i = w_0 + w_1 q_i. \tag{1}$$

w_1 stands for learnable parameters, w_0 stands for bias, and $q_i \in Q$.

Attribute Label Embeddings L**.** For each image, we retrieved a set of attribute label embeddings $L = \{l_1,l_2,\ldots,l_l\}$, $l_i \in \mathbb{R}^d$, with i ranging from one to the number of attributes. Attribute label embeddings were learned from an embedding layer of size $d \times l$. They represented the semantic information contained in the attribute label, that is, all possible attributes contained in the images.

3.3.2. Relation Exploitation Based on Transformer

Because the transformer architecture has shown great performance in capturing different and distant dependencies between variables in recent years [35,39,40], in this study, we utilized a transformer to model interactions between image features and attributes. We fed feature embeddings and attribute label embeddings into the transformer encoder simultaneously, and the attention mechanisms allowed the transformer to learn the dependencies between attributes and features.

Let $Z = \{z_1, z_2, \ldots, z_{h\times w}\}$, $z_i \in \mathbb{R}^d$, where Z represents the sum of the feature embeddings Q and position encoding P:

$$Z = Q + P. \tag{2}$$

Let $K = \{z_1, z_2, \ldots, z_{h\times w}, l_1, l_2, \ldots, l_l\}$ be the set of embeddings that are input to the transformer encoder(shown in Figure 3). In a transformer encoder, the weight of each embedding relative to other embeddings is learned through self-attention [34]. Let α_{ij} be the attention weight between embeddings $k_i \in K$ and $k_j \in K$. α_{ij} was computed using the following steps: First, we computed a normalized scalar attention coefficient α_{ij} between embeddings k_i and k_j as follows:

$$\alpha_{ij} = softmax((W^Q k_i)^\mathsf{T}(W^K k_j)/\sqrt{d}). \tag{3}$$

Then, each embedding k_i was updated to k_i' by calculating the weighted sum of all embeddings followed by a nonlinear ReLU layer:

$$k_i' = ReLU((\sum_{j=1}^{m} \alpha_{ij} W^V k_j) + b_1) + b_2. \tag{4}$$

Here, W^Q, W^K, and W^V were the query weight matrix, key weight matrix, and value weight matrix, respectively; b_1 and b_2 were bias vectors; m was equal to $h \times w + l$. This update procedure could be repeated for N layers, and the updated embeddings, k_i', were fed as inputs to the successive N transformer encoder layers. The learned weight matrices $\{W^Q, W^K, W^V\} \in \mathbb{R}^{d\times d}$ were not shared between layers. We denoted the final output of the transformer encoder after N layers as $K' = \{z_1', z_2', \ldots, z_{h\times w}', l_1', l_2', \ldots, l_l'\}$, where $L' = \{l_1', l_2', \ldots, l_l'\}$ was the attribute recognition outputs.

3.3.3. Attribute Recognition Classifier

After the features and attributes had been transferred via the transformer encoder, an independent feedforward network(FFN_i) was introduced to make the final label predictions. The FFN_i contained a single linear layer and its output was calculated as follows:

$$output_i = FFN_i(l'_i) = \sigma((w_i \cdot l'_i) + b_i). \tag{5}$$

where the weight w_i for label prediction $output_i$ is a vector $1 \times d$, b_i is a bias vector, and σ is a sigmoid function.

3.3.4. The Loss Function

We adopted the binary cross-entropy loss function [26] for safety attribute recognition using the following formula:

$$L(\hat{l}, l) = -\frac{1}{M}\sum_{m=1}^{M}(l^m \log(\sigma(\hat{l}^m)) + (1 - l^m)log(1 - \sigma(\hat{l}^m))). \tag{6}$$

where $\hat{l}$ and l represent the prediction results and the ground truth label, respectively, M is the number of attributes, and $\sigma(.)$ refers to the sigmoid activation function.

3.4. Fusion Processing Module

Owing to the limitations of the object detection algorithm, the worker detector might generate missed detections and false alarms in some frames, which may affect the accuracy of the safety attribute recognition. In our system, we adopted a simple voting method by combining the attribute recognition results with multiobject tracking trajectories, as shown in Figure 4. If it was not successfully detected and recognized, a judgment was given based on the previous 30 frames. If more than half of the previous 30 frames had detection and recognition events, the missed detection events were interpolated, and the tracking result was corrected. Otherwise, it was considered that the object disappeared. Finally, the detection and tracking results could be successfully smoothed. The successfully recognized or corrected results were transmitted to the interface for display.

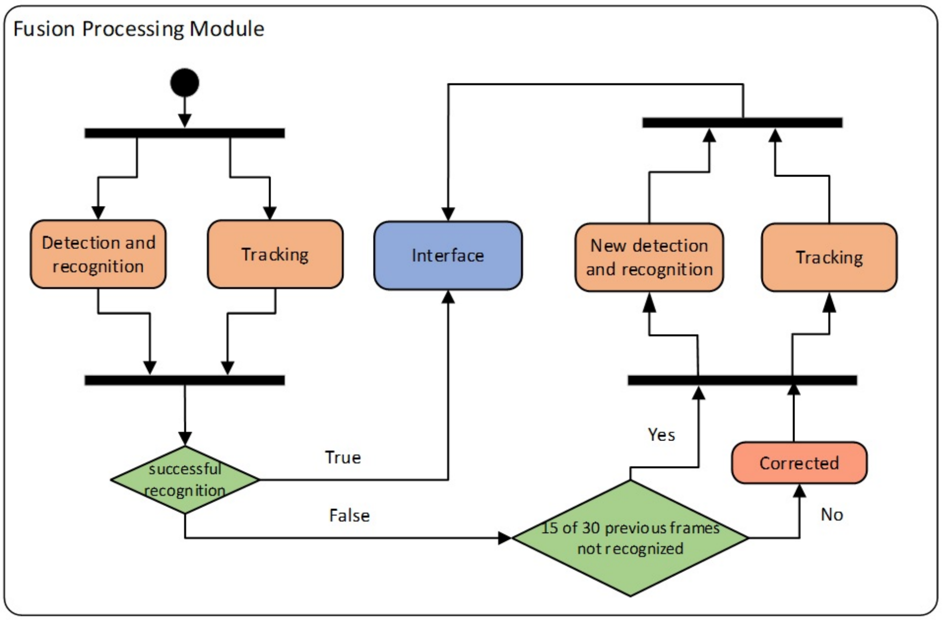


Figure 4. The fusion processing module. We fused the detection and recognition results on multiple-object-tracking trajectories.

4. Dataset and Experimental Details

4.1. Open-Site Monitoring Dataset

Because there are relatively few public datasets available for a detailed evaluation of safety helmet and harness monitoring systems in the community, we developed an open-site monitoring dataset for construction scenes containing three subdatasets: object detection, multiobject tracking, and attribute recognition. This benchmark dataset is crucial for evaluating safety helmet and harness monitoring systems in unconstrained environments.

(1) Object Detection Subdataset

The object detection subdataset can be used for safety helmet detection, safety harness detection, and worker detection in construction scenes. The images for safety helmet detection were selected from the public Safety Helmet Wearing Dataset (SHWD), which is an open-source dataset provided by Github (https://github.com/njvisionpower/Safety-Helmet-Wearing-Dataset, accessed on 1 January 2023).

As there is no public safety harness dataset available, we contributed a safety-harness-wearing subset containing 4196 images downloaded from Google or taken on construction sites. We manually labeled these data with LabelImg, and a VOC-format file was generated. We converted the subset from VOC format to txt format, as required by YOLOv5. The text-format file contained the annotation information of the images used for training or testing. Sample images from our proposed subdataset are shown in Figure 5.

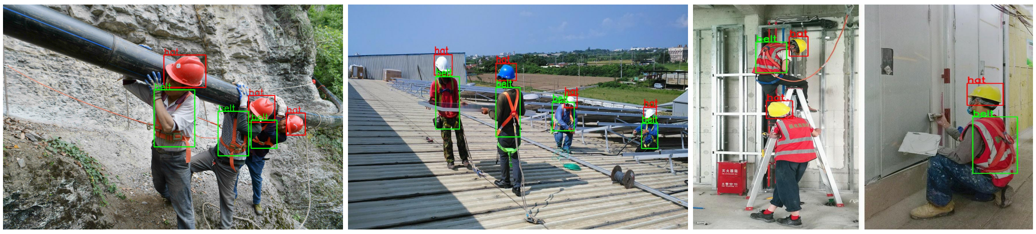


Figure 5. Sample images of our proposed object detection subdataset.

(2) Multiobject Tracking Subdataset

To evaluate our safety helmet and harness monitoring system for worker tracking, we contributed a multiobject tracking subdataset consisting of four videos with a total of 7200 frames, in which three videos were work at a height and the last one was work on the ground. All videos were annotated with the bounding boxes of workers, safety helmets, and harnesses, as well as the corresponding categories. The construction site videos were at 30 frames per second in a video of resolution 2560 × 1080. Sample images from the proposed video dataset are shown in Figure 6.

(3) Safety Attribute Recognition Subdataset

The safety attribute recognition subdataset of construction site workers was cropped from the proposed object detection subdataset and the multiobject tracking subdataset above. It consisted of 3633 images including multiple attributes such as age, gender, safety helmet, and safety harness. Among them, there were 164 images of workers wearing both helmets and harnesses, 2439 images of workers wearing safety helmets but no harnesses, 519 images of workers wearing safety harnesses but no helmets, and 511 images without safety helmets or harnesses. Sample examples are shown in Figure 7.

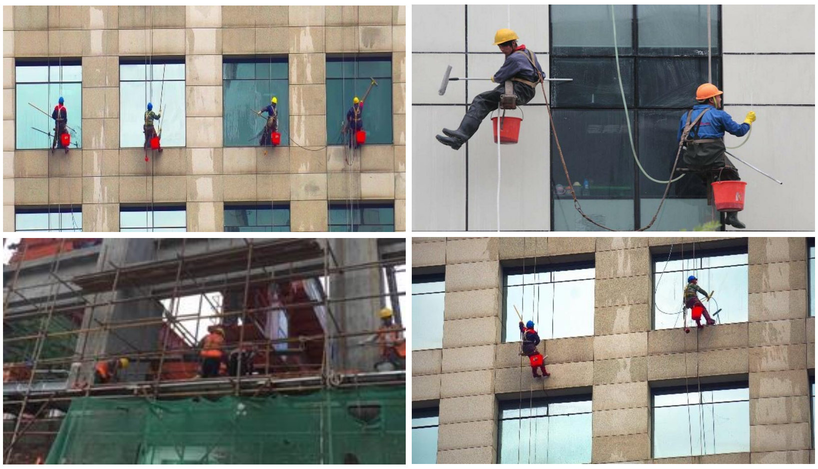


Figure 6. Sample images of our proposed multiobject tracking subdataset.

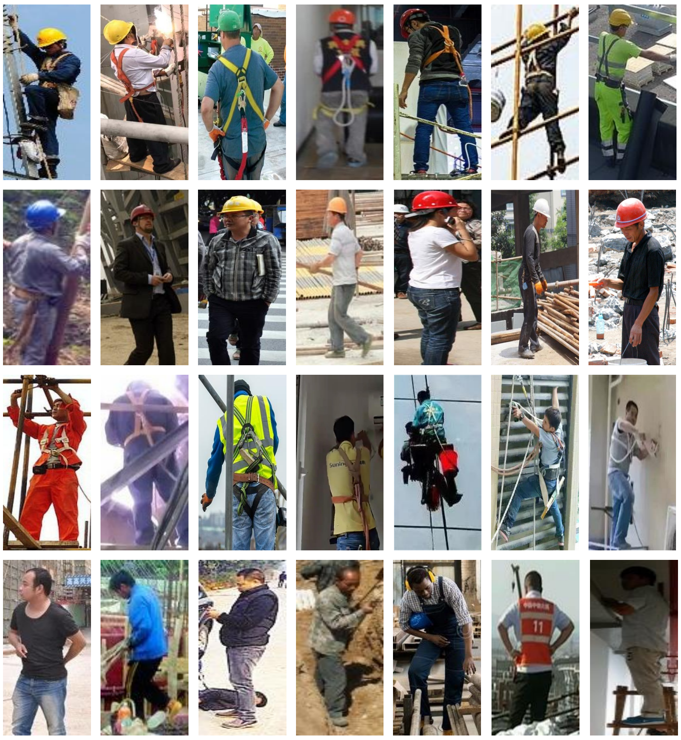


Figure 7. Sample images of our proposed attribute recognition subdataset. All images were normalized to 128 × 64. (**Top:**) Sample images of workers wearing safety helmets and harnesses. (**The second line:**) Samples of workers wearing safety helmets but no harnesses. (**The third line:**) Samples of workers wearing safety harnesses but no helmet. (**Bottom:**) Sample images without any safety helmet or harness.

4.2. Implementation Details

Image Feature Extractor. For fair comparisons, input images were resized to 224 × 224. Random horizontal mirroring, random rotation, and color jittering were used as data augmentation [46] during training. We used the ResNet50 [47] pretrained on ImageNet [48] as the backbone network to extract image features. We removed the last pooling layer and the full connection layer in the ResNet traditional network, and the output dimension was 2048, so we set the embedding size to $d = 2048$. Since the images were resized to 224 × 224, the output of ResNet50 was a $7 \times 7 \times d$ tensor. Therefore, there was a total of 49 feature embedding vectors.

Transformer Encoder. To allow a particular embedding to pay attention to multiple other embeddings (or multiple groups), our model used 4 attention heads [34]. We used a $L = 3$ layer transformer encoder with a residual layer [47] around each embedding update and layer norm [49].

Optimization. Our model was trained end-to-end. We used Adam [50] for the optimizer with betas = (0.9, 0.999) and a weight decay of 0. We trained the models with a batch size of 32 and a learning rate of 10^{-5}. We used dropout with $p = 0.1$ for the regularization.

5. Experimental Results

To verify the feasibility and accuracy of the proposed safety helmet and harness monitoring system, we conducted experiments using the following four aspects: object detection, object tracking, safety attribute recognition, and the visual interaction interface of the system.

5.1. Results and Analysis of Object Detection

We chose the YOLOv5s and YOLOv5x pretrained models on the COCO dataset [51] to train our proposed object detection subdataset. Precision, recall, and mean average precision (mAP) were adopted as evaluation metrics [52]. The relevant parameters, batch_size, and image_size of both models were set to 16 and 640 × 640, respectively, and 300 epochs were trained. Figure 8 shows the training results of our proposed dataset on the YOLOv5x weights. Table 1 shows the test results for each category of the two pretrained models.

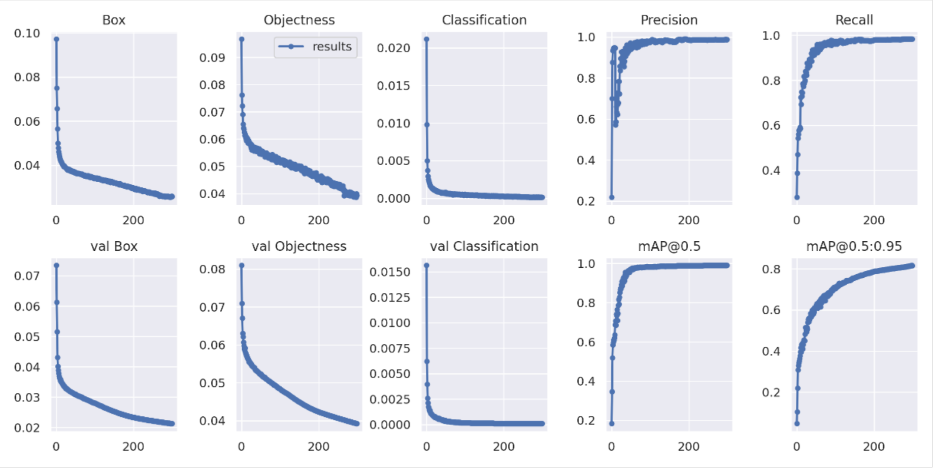


Figure 8. Training results of YOLOv5x on our proposed dataset.

Table 1 shows that the wearing detection of safety helmets and safety harnesses had a high precision, but a low recall. This meant that the detector generated a large number of missed detections, which affected the comprehensive metrics of mAP@.5 and mAP@.5:.95. Because of the limitation of the object detection in the task of safety helmet and harness monitoring, we transferred this task to a safety attribute recognition task. Please refer to Section 5.3 for the experimental evaluation of the safety attribute recognition.

Table 1. Detection results with different pretrained weights.

Network Model	Class	Precision	Recall	mAP@.5	mAP@.5:.95
YOLOv5s	All	0.957	0.875	0.921	0.652
	Head	0.959	0.897	0.948	0.684
	Safety helmet	0.969	0.933	0.976	0.742
	Safety harness	0.942	0.794	0.838	0.530
YOLOv5x	All	0.981	0.922	0.940	0.791
	Head	0.983	0.974	0.990	0.832
	Safety helmet	0.969	0.988	0.989	0.850
	Safety harness	0.993	0.806	0.841	0.693

5.2. Results and Analysis of Object Tracking

To evaluate object tracking in construction scenes, we selected the common Deep SORT algorithm to complete the worker tracking task, which was tested on the proposed multiobject tracking subdataset. We evaluated the task using the CLEAR metrics commonly used in multitarget tracking [53,54], including MOTA, FP, FN, and ID Sw. IDF1 presented in [55] evaluates different aspects of tracking performance, which evaluates the identity preservation ability and focuses more on the association performance. The model parameters were set as follows: Amax = 30 frames, confidence score = 0.4, and IOU threshold for NMS = 0.5. The test results are listed in Table 2. The execution speed of the system was approximately 25 fps, which met the real-time demand in real-world scenarios.

Table 2. The multiobject tracking results. ↑ indicates that higher scores are better, and ↓ means the opposite.

Detector	Tracker	MOTA↑	IDF1↑	MT↑	ML↓	ID Sw↓	FP↓	FN↓
YOLOv5s	Deep SORT	97.6%	98.9%	75.2%	12.2%	12	192	44
YOLOv5m		96.5%	98.3%	73.1%	13.9%	17	304	47
YOLOv5l		92.2%	96.5%	70.2%	16.1%	25	723	72
YOLOv5x		93.9%	97.3%	71.7%	15.4%	22	556	60

5.3. Results and Analysis of Safety Attribute Recognition

In this subsection, we compare the proposed safety attribute recognition method with some state-of-the-art methods on the proposed safety attribute recognition subdataset. According to previous works [24,25,32,56], we adopted five metrics to evaluate the attribute recognition performance, including a label-based metric called mean accuracy (mA), and four instance-based metrics including accuracy (Accu), precision (Prec), Recall, and F1 score. These metrics are widely used for pedestrian attribute recognition [10]. For a fair comparison, we report the performance of the proposed method based on the same settings.

The experimental results are listed in Table 3. It can be observed that the proposed method exhibited an improvement compared with the other methods. The mean accuracy was greater than 96%, which could be applied to actual scene applications.

Table 3. The attribute recognition results on our proposed safety attribute recognition subdataset. Best results are shown in bold.

Method	References	mA	Accuracy	Precision	Recall	F1 Score
Resnet50 [47]	CVPR'16	79.87	73.78	75.78	74.38	75.08
WRN [57]	CVPR'16	81.90	73.89	75.62	74.88	75.25
ALM [24]	ICCV'19	93.59	82.11	83.10	82.77	82.94
ViT [39]	ICLR'21	94.45	82.91	83.93	83.30	83.61
The proposed method		**96.44**	**86.76**	**87.47**	**88.38**	**87.34**

As shown in Figure 9, we visualized the localized attribute regions from two attributes, i.e., "Safety helmet" and "Safety harness". The proposed method helped to locate attribute-related regions for each attribute. For example, only the head region was considered when recognizing the attribute "Safety helmet" and the attention was stronger in the body region when recognizing the attribute "Safety harness". It was further proof that the proposed method could better model the relationship between attributes and feature regions by using self-attention and extracting more representative features.

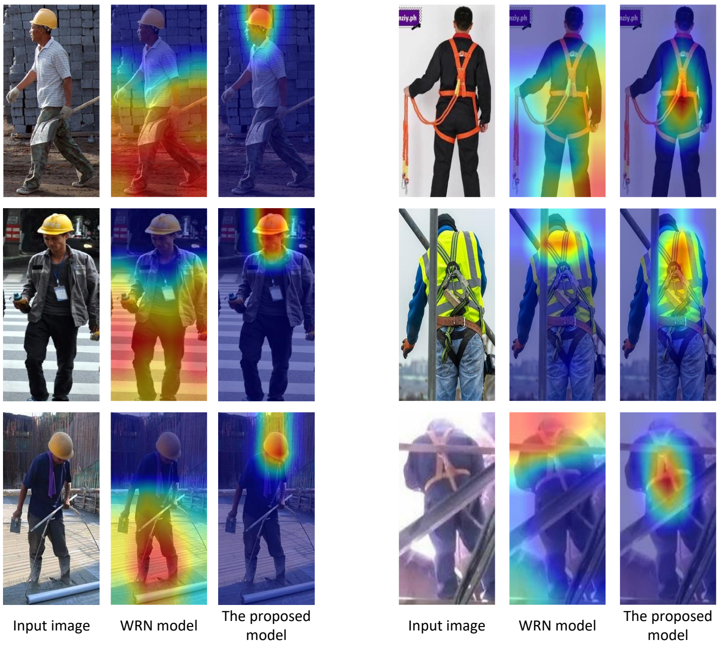
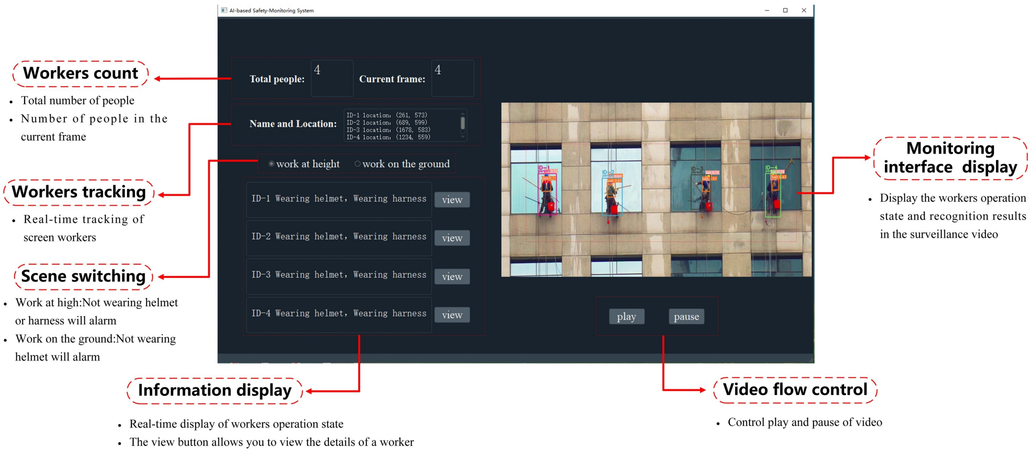


Figure 9. Attention regions of different models and attributes. The **left** is the visual feature map for the safety helmet attribute, and the **right** is the visual feature map for the safety harness attribute.

5.4. System Visual Interface Design and Display

For a better and more intuitive visual display of the recognition results, we developed a human–computer interaction interface for our system based on Pyqt5. The main interface had six functions, as shown in Figure 10: worker counting, worker tracking, scene switching, information display, video flow control, and monitoring screen display. Our system could work for two types of operating scenarios: working on the ground and working at height, depending on different rules and regulations at construction sites. Workers are allowed not to wear safety harnesses when working on the ground, but not wearing safety helmets causes an alarm. When working at a height, workers must wear safety helmets and safety harnesses, and the system will give an alarm if either of the two pieces of equipment is not worn. The UI is shown in Figure 11. We added the View button, which allowed users to see the details of a specific worker, as shown in Figure 12.



Total people: 4 Current frame: 4

Name and Location:
ID-1 location：(261, 573)
ID-2 location：(689, 599)
ID-3 location：(1678, 583)
ID-4 location：(1234, 559)

● work at height ○ work on the ground

ID-1 Wearing helmet，Wearing harness

ID-2 Wearing helmet，Wearing harness

ID-3 Wearing helmet，Wearing harness

ID-4 Wearing helmet，Wearing harness

Workers count
- Total number of people
- Number of people in the current frame

Workers tracking
- Real-time tracking of screen workers

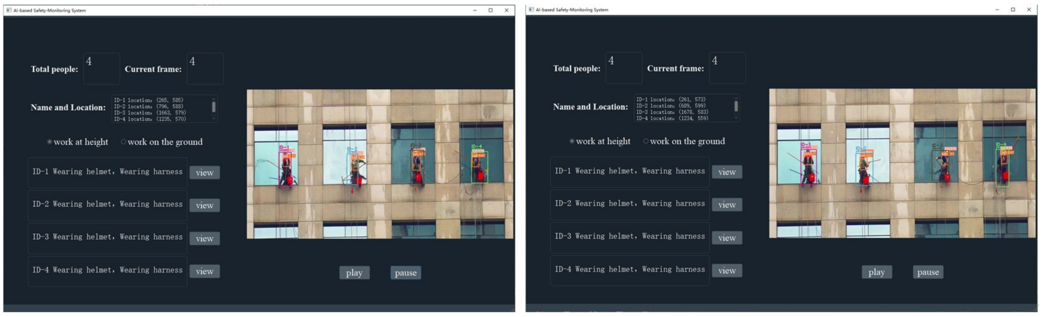
Scene switching
- Work at high:Not wearing helmet or harness will alarm
- Work on the ground:Not wearing helmet will alarm

Information display
- Real-time display of workers operation state
- The view button allows you to view the details of a worker

Video flow control
- Control play and pause of video

Monitoring interface display
- Display the workers operation state and recognition results in the surveillance video

Figure 10. The visual UI interface design.



Total people: 4 Current frame: 4

Name and Location:
ID-1 location: (265, 585)
ID-2 location: (796, 588)
ID-3 location: (1663, 579)
ID-4 location: (1235, 570)

● work at height ○ work on the ground

ID-1 Wearing helmet, Wearing harness
ID-2 Wearing helmet, Wearing harness
ID-3 Wearing helmet, Wearing harness
ID-4 Wearing helmet, Wearing harness

Total people: 4 Current frame: 4

Name and Location:
ID-1 location: (261, 573)
ID-2 location: (689, 599)
ID-3 location: (1678, 583)
ID-4 location: (1234, 559)

● work at height ○ work on the ground

ID-1 Wearing helmet, Wearing harness
ID-2 Wearing helmet, Wearing harness
ID-3 Wearing helmet, Wearing harness
ID-4 Wearing helmet, Wearing harness

Figure 11. The visual UI interface display.

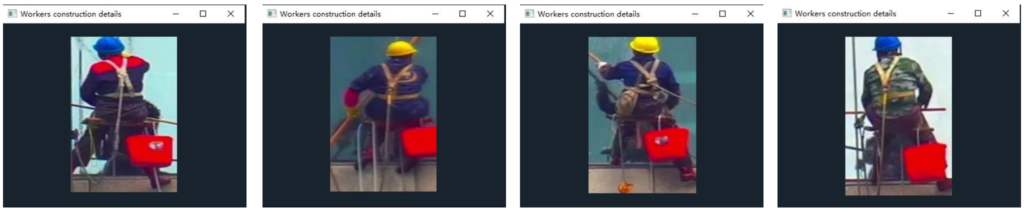

Workers construction details
Workers construction details
Workers construction details
Workers construction details

Figure 12. The worker's construction details.

6. Discussion

6.1. The Superiority of Attribute Recognition for Safety Helmet and Harness Monitoring

A safety helmet and harness monitoring system mainly refers to the use of video technology to monitor a construction scene and record the scene images in real time. With the development of technology, traditional safety helmet and harness monitoring systems have been unable to meet the needs of construction scenarios with frequent accidents. People are paying more and more attention to the practical application of artificial intelligence technology in safety helmet and harness monitoring systems. The current CV-based methods are all based on object detection technology, which has the defect of an insufficient accuracy due to occlusions or small targets. We proposed to transform this problem into an image's semantic attribute recognition problem and compared the difference between the two as shown in Figure 13. It can be seen that the general object detection model YOLOv5 is prone to missed detection for occlusion situations; for example, the harnesses are often missed due to occlusion. Our recognition model based on attribute knowledge modeling can utilize semantic information to accurately identify occluded attributes. Our model learns the association between features and attributes, which alleviates the impact of occlusion in practical applications. Moreover, our attribute recognition framework can recognize scene information. Scene information is useful for practical applications, for example, when working on the ground, where safety harnesses are not required. However, scene information is also an indispensable part of a safety helmet and harness monitoring system. Our proposed remote-vision-based safety helmet and harness monitoring system framework addresses these deficiencies well. It is not only more robust to occluded situations but also can recognize unseen classes such as scene information.

For a quantitative analysis, to better demonstrate the superiority of our framework, we trained the attribute recognition model (ours) and the general object detection model (YOLOv5) under the same settings (dataset, optimizer, etc.). We use precision and recall for the evaluation and the results are shown in Figure 14. It can be seen that our proposed method based on attribute recognition was superior over the method based on object detection.

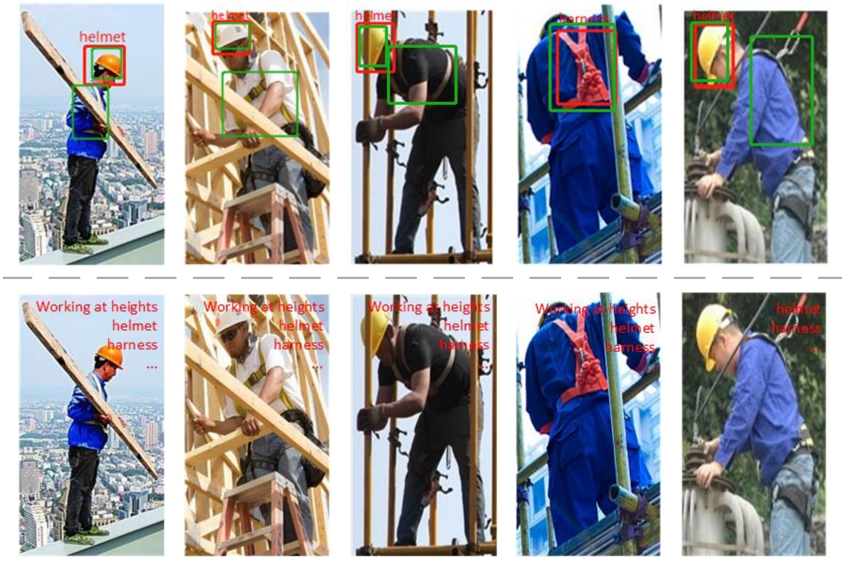


Figure 13. The difference between the object detection model and the attribute recognition model. The top row is the processing results of the general object detection models (YOLOv5), and the bottom row is the processing results of our proposed attribute recognition model (ours). In object detection, green bounding boxes represent ground truth, and red boxes represent predicted results. In attribute recognition, the result of the recognition is displayed directly on the image in the form of text.

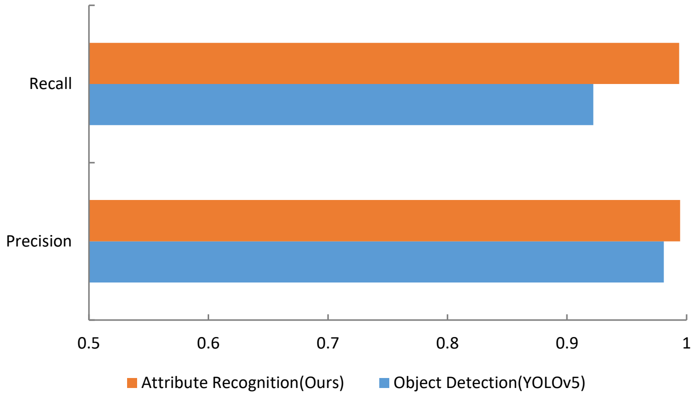


Figure 14. Comparison of recall and precision between attribute recognition model (ours) and object detection model (YOLOv5) under the same dataset.

6.2. The Effectiveness of Transformer on Attribute Knowledge Modeling

Some methods [40,43,58,59] have shown that the transformer architecture can better capture the relationship between visual features and process sequences in parallel during training. In this paper, we presented a novel attribute knowledge modeling network based on the transformer architecture. We aimed to improve recognition performance by exploiting self-supervision and self-attention mechanisms to explore the relations between attributes and image features fully.

In a quantitative analysis, we drew the mean accuracy (mA) results of the ResNet50 and the proposed model as shown in Figure 15. We can see that the performance of the model was greatly improved after adding the transformer.

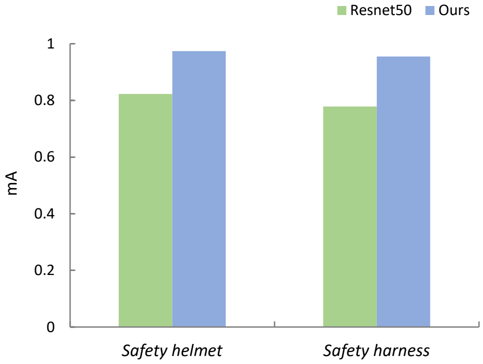


Figure 15. The mA comparison of different models. Resnet50 stands for backbone. Ours stands for adding a transformer after Resnet50 to learn relational knowledge.

As a qualitative analysis, Figure 16 shows a comparison of the attention maps of Resnet50 [47] and the proposed transformer-based method (ours) for different attributes. It can be seen that the transformer-based model focused more precisely on the regions that needed attention. It reduced the entangled mapping relationship between different regions and improved the robustness and accuracy of the mapping relations between regions and attributes.

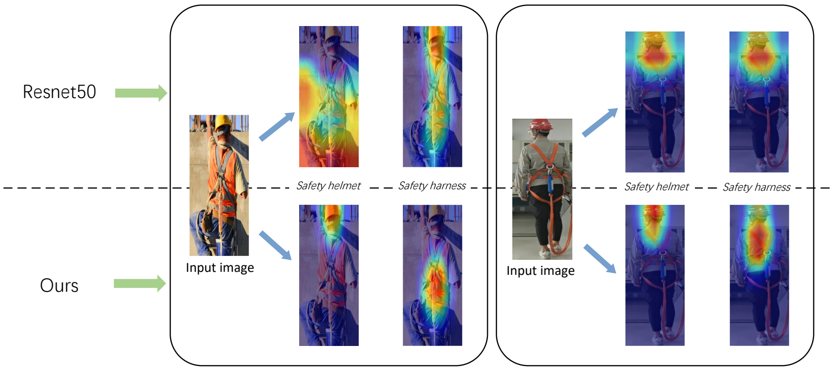


Figure 16. Attention comparison of different models. Visualization of different spatial information extracted by the Resnet50 model (the **top** part) and our method (the **bottom** part).

6.3. The Significance of the Open-Site Monitoring Dataset

As far as we know, our proposed open-site monitoring dataset is the first formally proposed dataset for construction scenarios in the industrial field. Although there are currently a few public datasets in the community, such as for the detection of safety helmets, these datasets do not have a clear, concrete, and unified setting. Most of the current datasets are collected in real life and randomly split. This results in a large number of identical pedestrian identities in the training and test set with the same image features. As a result, the existing datasets' settings are inconsistent with real-world applications.

Given the problems of existing datasets, the reasons why an open-site monitoring dataset is crucial for industrial applications and academic research are given as follows. First, because there are relatively few public datasets available for a detailed evaluation of safety helmet and harness monitoring systems in the community, we developed an open-site monitoring dataset for construction scenes containing three subdatasets: object detection, multiobject tracking, and attribute recognition. Second, whether used as a primary task in video surveillance or an auxiliary task in person retrieval, pedestrian identities of the test set barely overlap with the identities of the training set. Finally, we provided a strong transformer-based baseline on this dataset for follow-up studies.

7. Conclusions

In this study, we developed a novel and practical safety helmet and harness monitoring system to identify whether workers wear safety helmets and harnesses. Because the object detection algorithm may not be sufficiently accurate when applied in construction scenes, we proposed to transfer the object detection problem into a recognition problem of the semantic attributes of images. To make the identification more accurate, we proposed a novel end-to-end framework for safety attribute recognition that made full use of the spatial and semantic relations between images and attributes. Specifically, this study attempted to introduce the transformer into an attribute recognition task and achieved improved performance in attribute recognition. We contributed a novel open-site construction scene dataset that included three subdatasets for object detection, object tracking, and attribute recognition under construction site scenarios. Finally, the experimental results demonstrated the effectiveness and efficiency of this remote-vision-based safety helmet and harness monitoring system.

Application Scope and Limitation. Compared with some current cutting-edge technologies, our security monitoring has the advantages of a low cost, an easy deployment, and being more intuitive. However, we used a network of video surveillance cameras on construction sites. If workers are in danger of falling in places that cannot be monitored, e.g., high-rise building construction, video streaming data cannot be collected. This is a limitation of our system. Recently, Shanti et al. [19] proposed utilizing drones to monitor workers at heights in real time, which is a great strategy. In future work, we will consider how to optimize this problem, such as using a computer-vision-based drone, which is very beneficial to the practical application of our system.

Author Contributions: Conceptualization, S.Z. and S.M.; methodology, X.W. and Y.L.; software, J.L.; validation, X.W., Y.L. and J.L.; writing—original draft preparation, X.W., S.Z., S.W. and S.M.; supervision, S.W. All authors have read and agreed to the published version of the manuscript.

Funding: The work is supported by National Natural Science Foundation of China (no. 62271409 and 62171381).

Data Availability Statement: Data available on request from the authors.

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

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional neural network
SSD	Single-shot multibox detector
MTL	Multitask learning
GCN	Graph convolutional networks
LSTM	Long short-term memory
HPSS	Hierarchical positive sample selection
MDA	Multidirectional attention
OSHA	Occupational Safety and Health Administration
PPE	Personal protective equipment

References

1. Jeong, G.; Kim, H.; Lee, H.S.; Park, M.; Hyun, H. Analysis of safety risk factors of modular construction to identify accident trends. *J. Asian Archit. Build. Eng.* **2022**, *21*, 1040–1052. [CrossRef]
2. OSHA. Available online: https://www.osha.gov/Publications/OSHA3252/3252.html (accessed on 6 July 2019).
3. Mei, S.; Geng, Y.; Hou, J.; Du, Q. Learning hyperspectral images from RGB images via a coarse-to-fine CNN. *Sci. China Inf. Sci.* **2022**, *65*, 1–14. [CrossRef]
4. Han, G.; Zhu, M.; Zhao, X.; Gao, H. Method based on the cross-layer attention mechanism and multiscale perception for safety helmet-wearing detection. *Comput. Electr. Eng.* **2021**, *95*, 107458. [CrossRef]
5. Li, Z.; Xie, W.; Zhang, L.; Lu, S.; Xie, L.; Su, H.; Du, W.; Hou, W. Toward Efficient Safety Helmet Detection Based on YoloV5 With Hierarchical Positive Sample Selection and Box Density Filtering. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–14. [CrossRef]

6. 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]
7. Wu, J.; Cai, N.; Chen, W.; Wang, H.; Wang, G. Automatic detection of hardhats worn by construction personnel: A deep learning approach and benchmark dataset. *Autom. Constr.* **2019**, *106*, 102894. [CrossRef]
8. Fang, C.; Xiang, H.; Leng, C.; Chen, J.; Yu, Q. Research on Real-Time Detection of Safety Harness Wearing of Workshop Personnel Based on YOLOv5 and OpenPose. *Sustainability* **2022**, *14*, 5872. [CrossRef]
9. Fang, W.; Ding, L.; Luo, H.; Love, P.E. Falls from heights: A computer vision-based approach for safety harness detection. *Autom. Constr.* **2018**, *91*, 53–61. [CrossRef]
10. Wang, X.; Zheng, S.; Yang, R.; Zheng, A.; Chen, Z.; Tang, J.; Luo, B. Pedestrian attribute recognition: A survey. *Pattern Recognit.* **2022**, *121*, 108220. [CrossRef]
11. Ray, S.J.; Teizer, J. Real-time construction worker posture analysis for ergonomics training. *Adv. Eng. Inform.* **2012**, *26*, 439–455. [CrossRef]
12. Seo, J.; Han, S.; Lee, S.; Kim, H. Computer vision techniques for construction safety and health monitoring. *Adv. Eng. Inform.* **2015**, *29*, 239–251. [CrossRef]
13. Yan, X.; Li, H.; Li, A.R.; Zhang, H. Wearable IMU-based real-time motion warning system for construction workers' musculoskeletal disorders prevention. *Autom. Constr.* **2017**, *74*, 2–11. [CrossRef]
14. Wonil, L.; Seto, E.; Lin, K.; Migliaccio, G. An evaluation of wearable sensor s and their placements for analyzing construction worker's trunk posture i n laboratory conditions. *Appl. Erg.* **2017**, *65*, 424–436.
15. Kolar, Z.; Chen, H.; Luo, X. Transfer learning and deep convolutional neural networks for safety guardrail detection in 2D images. *Autom. Constr.* **2018**, *89*, 58–70. [CrossRef]
16. Fang, Q.; Li, H.; Luo, X.; Ding, L.; Luo, H.; Rose, T.M.; An, W. Detecting non-hardhat-use by a deep learning method from far-field surveillance videos. *Autom. Constr.* **2018**, *85*, 1–9. [CrossRef]
17. Shanti, M.Z.; Cho, C.S.; Byon, Y.J.; Yeun, C.Y.; Kim, T.Y.; Kim, S.K.; Altunaiji, A. A novel implementation of an ai-based smart construction safety inspection protocol in the uae. *IEEE Access* **2021**, *9*, 166603–166616. [CrossRef]
18. Alrayes, F.S.; Alotaibi, S.S.; Alissa, K.A.; Maashi, M.; Alhogail, A.; Alotaibi, N.; Mohsen, H.; Motwakel, A. Artificial Intelligence-Based Secure Communication and Classification for Drone-Enabled Emergency Monitoring Systems. *Drones* **2022**, *6*, 222. [CrossRef]
19. Shanti, M.Z.; Cho, C.S.; de Soto, B.G.; Byon, Y.J.; Yeun, C.Y.; Kim, T.Y. Real-time monitoring of work-at-height safety hazards in construction sites using drones and deep learning. *J. Saf. Res.* **2022**, *83*, 364–370. [CrossRef]
20. Zhu, J.; Liao, S.; Lei, Z.; Yi, D.; Li, S. Pedestrian attribute classification in surveillance: Database and evaluation. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Sydney, NSW, Australia, 2–8 December 2013; pp. 331–338.
21. Deng, Y.; Luo, P.; Loy, C.C.; Tang, X. Pedestrian attribute recognition at far distance. In Proceedings of the 22nd ACM International Conference on Multimedia, New York, NY, USA, 3–7 November 2014; pp. 789–792.
22. Zhao, X.; Sang, L.; Ding, G.; Han, J.; Di, N.; Yan, C. Recurrent attention model for pedestrian attribute recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 9275–9282.
23. Tan, Z.; Yang, Y.; Wan, J.; Hang, H.; Guo, G.; Li, S.Z. Attention-based pedestrian attribute analysis. *IEEE Trans. Image Process.* **2019**, *28*, 6126–6140. [CrossRef]
24. Tang, C.; Sheng, L.; Zhang, Z.; Hu, X. Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4997–5006.
25. Liu, X.; Zhao, H.; Tian, M.; Sheng, L.; Shao, J.; Yi, S.; Yan, J.; Wang, X. Hydraplus-net: Attentive deep features for pedestrian analysis. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 350–359.
26. Li, D.; Chen, X.; Huang, K. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In Proceedings of the 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), Kuala Lumpur, Malaysia, 3–6 November 2015; pp. 111–115.
27. Mei, S.; Li, X.; Liu, X.; Cai, H.; Du, Q. Hyperspectral image classification using attention-based bidirectional long short-term memory network. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5509612. [CrossRef]
28. Mei, S.; Chen, X.; Zhang, Y.; Li, J.; Plaza, A. Accelerating convolutional neural network-based hyperspectral image classification by step activation quantization. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5502012. [CrossRef]
29. Sarfraz, M.S.; Schumann, A.; Wang, Y.; Stiefelhagen, R. Deep view-sensitive pedestrian attribute inference in an end-to-end model. *arXiv* **2017**, arXiv:1707.06089.
30. Wang, J.; Zhu, X.; Gong, S.; Li, W. Attribute recognition by joint recurrent learning of context and correlation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 531–540.
31. Zhao, X.; Sang, L.; Ding, G.; Guo, Y.; Jin, X. Grouping attribute recognition for pedestrian with joint recurrent learning. In Proceedings of the IJCAI, Stockholm, Sweden, 13–19 July 2018; Volume 2018, p. 27.
32. Li, Q.; Zhao, X.; He, R.; Huang, K. Visual-semantic graph reasoning for pedestrian attribute recognition. In Proceedings of the AAAI conference on artificial intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8634–8641.
33. Tan, Z.; Yang, Y.; Wan, J.; Guo, G.; Li, S.Z. Relation-aware pedestrian attribute recognition with graph convolutional networks. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12055–12062.

34. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. [CrossRef]
35. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
36. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv* **2019**, arXiv:1909.11942.
37. Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q.V.; Salakhutdinov, R. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv* **2019**, arXiv:1901.02860.
38. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
39. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
40. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 213–229.
41. He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; Jiang, W. Transreid: Transformer-based object re-identification. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual Conference, 10–17 October 2021; pp. 15013–15022.
42. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in vision: A survey. *ACM Comput. Surv. (CSUR)* **2022**, *54*, 1–41. [CrossRef]
43. Gabeur, V.; Sun, C.; Alahari, K.; Schmid, C. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 214–229.
44. Cornia, M.; Stefanini, M.; Baraldi, L.; Cucchiara, R. Meshed-memory transformer for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10578–10587.
45. Chen, S.; Hong, Z.; Liu, Y.; Xie, G.S.; Sun, B.; Li, H.; Peng, Q.; Lu, K.; You, X. Transzero: Attribute-guided transformer for zero-shot learning. In Proceedings of the AAAI, Virtual Conference, 22 February–1 March 2022; Volume 2, p. 3.
46. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 1–48. [CrossRef]
47. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
48. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
49. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
50. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
51. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 740–755.
52. Padilla, R.; Netto, S.L.; Da Silva, E.A. A survey on performance metrics for object-detection algorithms. In Proceedings of the 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), Niteroi, Brazil, 1–3 July 2020; pp. 237–242.
53. Zhang, S.; Wang, J.; Wang, Z.; Gong, Y.; Liu, Y. Multi-target tracking by learning local-to-global trajectory models. *PR* **2015**, *48*, 580–590. [CrossRef]
54. Zhang, S.; Huang, J.B.; Lim, J.; Gong, Y.; Wang, J.; Ahuja, N.; Yang, M.H. Tracking persons-of-interest via unsupervised representation adaptation. *Int. J. Comput. Vis.* **2020**, *128*, 96–120. [CrossRef]
55. Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; Tomasi, C. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 17–35.
56. Li, D.; Zhang, Z.; Chen, X.; Huang, K. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. *IEEE Trans. Image Process.* **2018**, *28*, 1575–1590. [CrossRef] [PubMed]
57. Zagoruyko, S.; Komodakis, N. Wide residual networks. *arXiv* **2016**, arXiv:1605.07146.
58. Huang, L.; Wang, W.; Chen, J.; Wei, X.Y. Attention on attention for image captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Long Beach, CA, USA, 16–20 June 2019; pp. 4634–4643.
59. Pan, Y.; Yao, T.; Li, Y.; Mei, T. X-linear attention networks for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual Conference, 14–19 June 2020; pp. 10971–10980.

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