CVPDL: Computer Vision Practice with Deep Learning

HW #1 Object Detection

生機碩二 許喬淇 R12631001

1. Architecture of Real-Time DEtection TRansformer (RT-DETR)

RT-DETR consists of a backbone, an efficient hybrid encoder, and a Transformer decoder with auxiliary prediction heads (Zhao, 2024) (Figure 1-1). The overview of RT-DETR is illustrated in Figure 4. Specifically, we feed the features from the last three stages of the backbone {S3, S4, S5} into the encoder. The efficient hybrid encoder transforms multi-scale features into a sequence of image features through intra-scale feature interaction and cross-scale feature fusion (Figure 1-2). Subsequently, the uncertainty-minimal query selection is employed to select a fixed number of encoder features to serve as initial object queries for the decode. Finally, the decoder with auxiliary prediction heads iteratively optimizes object queries to generate categories and boxes.

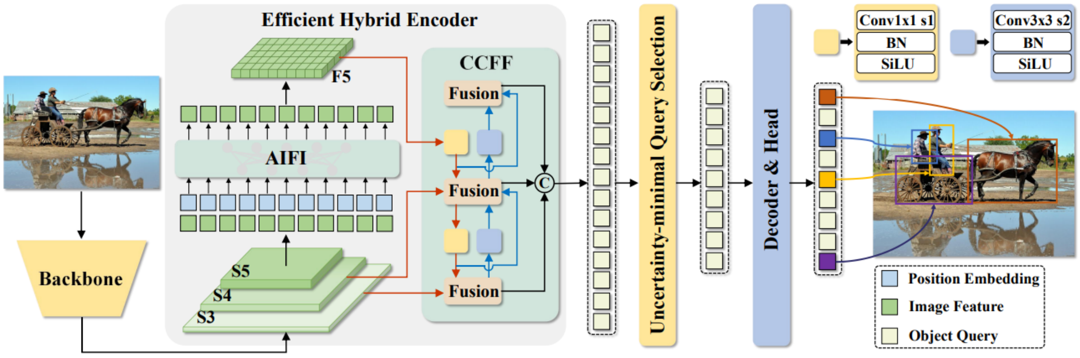


Figure 1-1 Overview of RT-DETR.

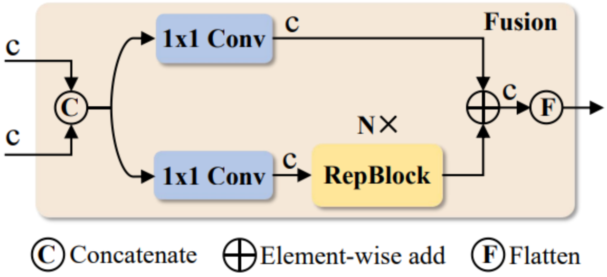


Figure 1-2 The fusion block in CCFF.

2. Training Detail

2.1. Train Setting

- Device: NVIDIA GeForce RTX 3090, 24030MiB *2
- Model: Based on the author's experimental results on the COCO val2017 dataset, the best-performing RT-DETR Extra-Large model (RT-DETR-X) was selected, which achieved an AP of 54.8%.
- Pretrained model: rtdetr-x.pt (can be download at https://docs.ultralytics.com/models/rtdetr/#usage-examples)
- Epochs: 300
- Patience: 50
- Batch: 8
- Input image size: 960 (resize image)
- Device: "0,1"
- Optimizer: Adam
- Initial learning rate: 0.00015
- Final learning rate: 0.00015
- Warmup_epochs: 3
- Weight of the classification loss: 0.6
- Weight of the box loss: 7.5
- Val: True (Enables validation during training)
- Augment=True

2.2. Augmentation Method

- Adjusts the hue of the image
- Alters the saturation of the image
- Modifies the brightness of the image
- Translates the image horizontally and vertically
- Scales the image

- Flips the image horizontally (left to right)
- Mosaics the image (combines four training images into one)
- Erases parts of the image (randomly erases a portion of the image).

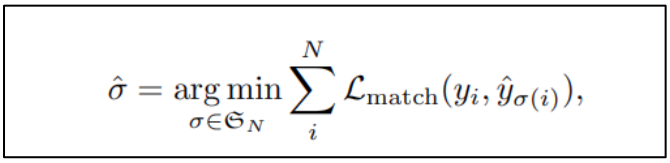
2.3. Loss function

Object detection set prediction loss: loss produces an optimal bipartite matching between predicted and ground truth objects, and then optimize object-specific (bounding box) losses.

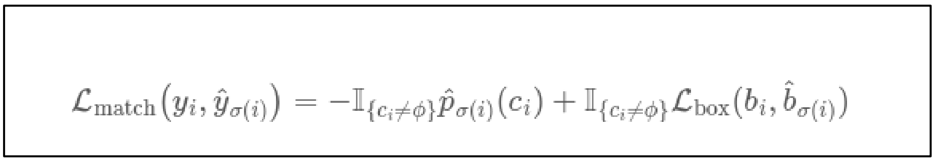
y : ground truth set of objects,

$\hat{y}$: set of *N* predictions. (*N* >=number of ground truth objects)

Consider y also as a set of size *N* padded with Ø

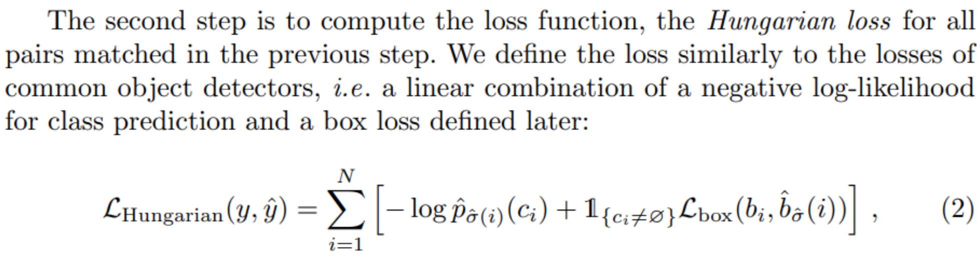


$$\hat{\sigma} = \underset{\sigma\in\mathfrak{S}_N}{\arg\min} \sum_{i}^{N} \mathcal{L}_{\mathrm{match}}(y_i, \hat{y}_{\sigma(i)}),$$



$$\mathcal{L}_{\mathrm{match}}\left(y_i, \hat{y}_{\sigma(i)}\right) = -\mathbb{I}_{\{c_i\neq\phi\}}\hat{p}_{\sigma(i)}(c_i) + \mathbb{I}_{\{c_i\neq\phi\}}\mathcal{L}_{\mathrm{box}}(b_i, \hat{b}_{\sigma(i)})$$

$L_{match}(y_i\ ,\hat{y}_{\sigma(i)})$ is a pair-wise matching cost between ground truth y_i and a prediction with index $\sigma(i)$.



The second step is to compute the loss function, the *Hungarian loss* for all pairs matched in the previous step. We define the loss similarly to the losses of common object detectors, *i.e.* a linear combination of a negative log-likelihood for class prediction and a box loss defined later:

$$\mathcal{L}_{\mathrm{Hungarian}}(y, \hat{y}) = \sum_{i=1}^{N} \left[-\log \hat{p}_{\hat{\sigma}(i)}(c_i) + \mathbb{1}_{\{c_i\neq\varnothing\}}\mathcal{L}_{\mathrm{box}}(b_i, \hat{b}_{\hat{\sigma}}(i)) \right] , \tag{2}$$

Specifically, the feature uncertainty $\mathcal{U}$ is defined as the discrepancy between the predicted distributions of localization $\mathcal{P}$ and classification $\mathcal{C}$ in Eq. (2). To minimize the uncertainty of the queries, we integrate the uncertainty into the loss function for the gradient-based optimization in Eq. (3).

$$\mathcal{U}(\hat{\boldsymbol{\mathcal{X}}}) = \|\mathcal{P}(\hat{\boldsymbol{\mathcal{X}}}) - \mathcal{C}(\hat{\boldsymbol{\mathcal{X}}})\|, \hat{\boldsymbol{\mathcal{X}}} \in \mathbb{R}^D \tag{2}$$

$$\mathcal{L}(\hat{\boldsymbol{\mathcal{X}}}, \hat{\boldsymbol{\mathcal{Y}}}, \boldsymbol{\mathcal{Y}}) = \mathcal{L}_{box}(\hat{\mathbf{b}}, \mathbf{b}) + \mathcal{L}_{cls}(\mathcal{U}(\hat{\boldsymbol{\mathcal{X}}}), \hat{\mathbf{c}}, \mathbf{c}) \tag{3}$$

where $\hat{\boldsymbol{\mathcal{Y}}}$ and $\boldsymbol{\mathcal{Y}}$ denote the prediction and ground truth, $\hat{\boldsymbol{\mathcal{Y}}} = \{\hat{\mathbf{c}}, \hat{\mathbf{b}}\}$, $\hat{\mathbf{c}}$ and $\hat{\mathbf{b}}$ represent the category and bounding box respectively, $\hat{\boldsymbol{\mathcal{X}}}$ represent the encoder feature.

3. Performance for validation set

3.1. Evaluation Setting

- Confident threshold = 0.55
- Input image size = 960
- IoU threshold = 0.5

3.2. Performance

The trained model achieved mAP50-95 scores of 0.6674 and 0.4515 using cvpdl and RT-DETR evaluation metrics, respectively (Table 3-1).

Metrics	mAP50	mAP75	mAP50-95
RT-DETR	0.6334	0.5074	0.4515
CVPDL	0.8496	0.7435	**0.6674**

Table 3-1 Model performance by using RT-DETR and CVPDL metrics.

Table 3-2 Model performance for each category by using RT-DETR metrics.

Category	mAP50	mAP75	mAP50-95
Person	0.877	0.731	0.73
Ear	0.861	0.598	0.598
Earmuffs	0.561	0.348	0.348
Face	0.925	0.726	0.726
Face-guard	0.462	0.310	0.31
Face-mask-medical	0.750	0.541	0.541
Foot	0.340	0.187	0.187
Tools	0.382	0.236	0.236
Glasses	0.727	0.469	0.469
Gloves	0.590	0.406	0.406
Helmet	0.650	0.467	0.467
Hands	0.852	0.627	0.627
Head	0.900	0.738	0.739
Medical-suit	0.290	0.184	0.184
Shoes	0.658	0.427	0.427
Safety-suit	0.410	0.307	0.307
Safety-vest	0.520	0.371	0.371
All	**0.633**	**0.507**	**0.452**

4. Visualization and discussion

The Long Tail Effect is a concept that describes a specific market or business model (Figure 4-1). It refers to the phenomenon in a market where a small number of popular products (the head) coexist with a large number of niche products (the long tail).

Key characteristics:

- Head - a few popular items account for the majority of sales.
- Long tail - a large number of niche products, while individually low in sales, collectively form a significant total.

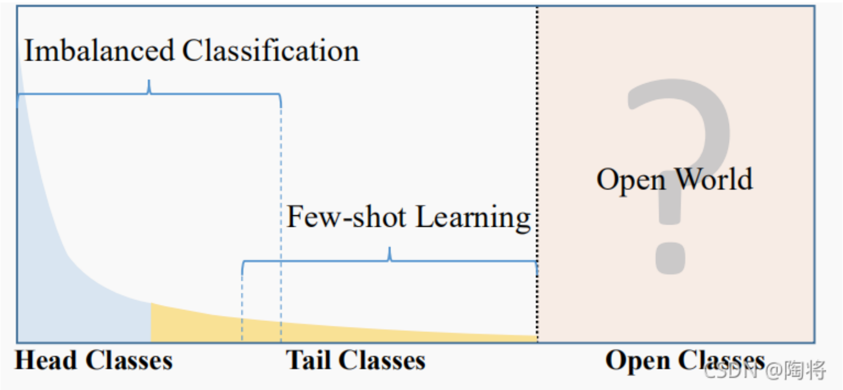


Figure 4-1 Long tail effect (https://blog.csdn.net/weixin_42111770/article/details/120421812).

Regarding model performance, categories with more objects, such as hands, person, head, face, and ear, generally achieved better mAP50-95 scores (Figure 4-2). Conversely, categories with fewer training objects, like safety-vest, earmuffs, safety-suit, medical-suit, and face-guard, showed poorer performance. However, some categories deviate from the Long Tail Effect. For instance, while the 'hands' class has the largest number of instances, it doesn't achieve the highest performance. Similarly, the 'tools' class, despite having 2,469 instances, ranks third from the bottom in terms of mAP. These differences can be explained by the unique features of each class. In the ground truth data, hands are often annotated without clear shapes (Figure 4-3), which may lead to the unexpected performance results. As for the 'tools' category, it includes many different types of items (Figure 4-4), which might confuse the model and result in undetected objects (Figure 4-5).

Additionally, we discovered that the model can detect blurry and dark objects. However, it does have some limitations. For instance, the model failed to detect a person seated in a device (Figure 4-6). Similarly, shoes that were only partially visible in the frame were not detected. Furthermore, the model struggles to detect objects that are too small.

Regarding successful predictions, the model demonstrates a strong ability to detect partial faces (Figure 4-7). Even when only the mouth is visible, the model can still accurately identify and detect the face.

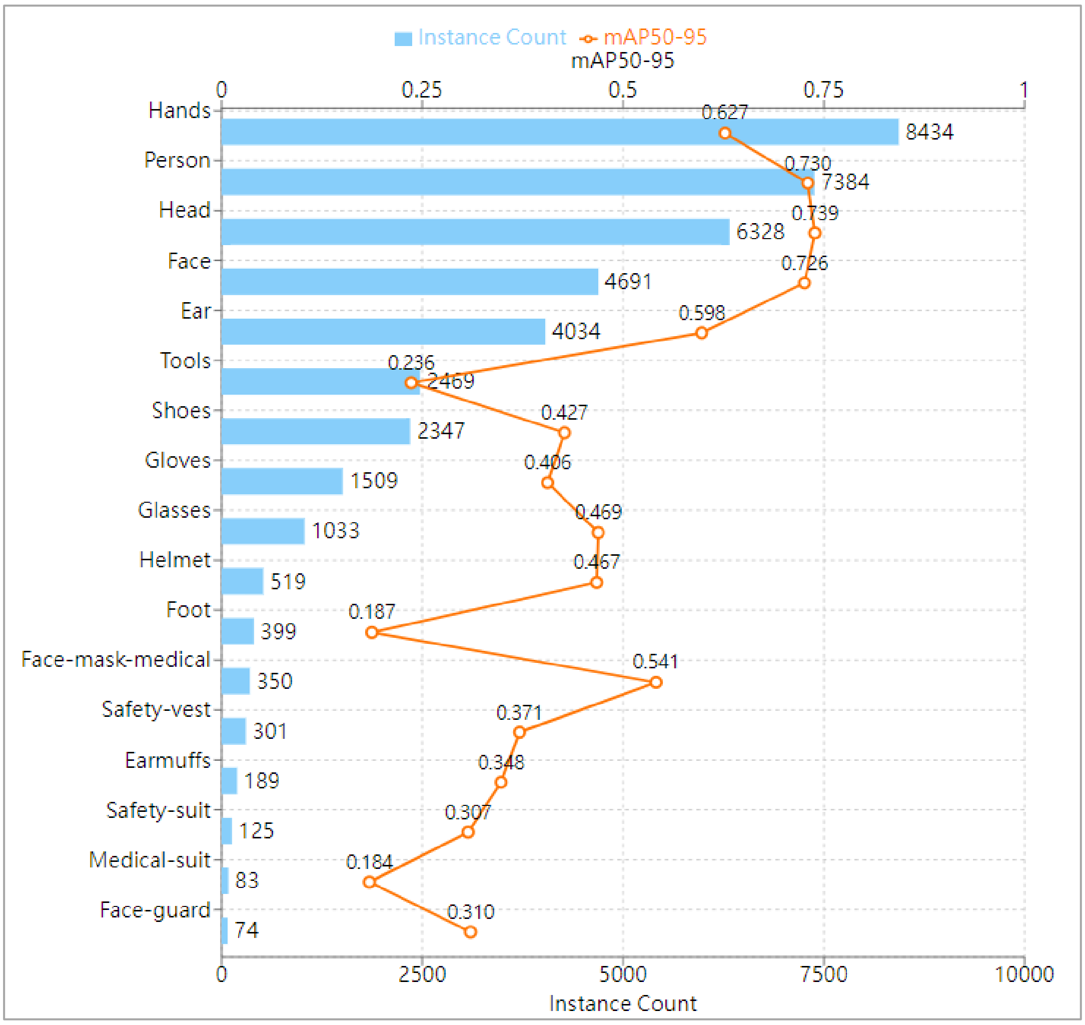


Figure 4-2 Dataset distribution with mAP50-95 using RT-DETR metrics for all categories.

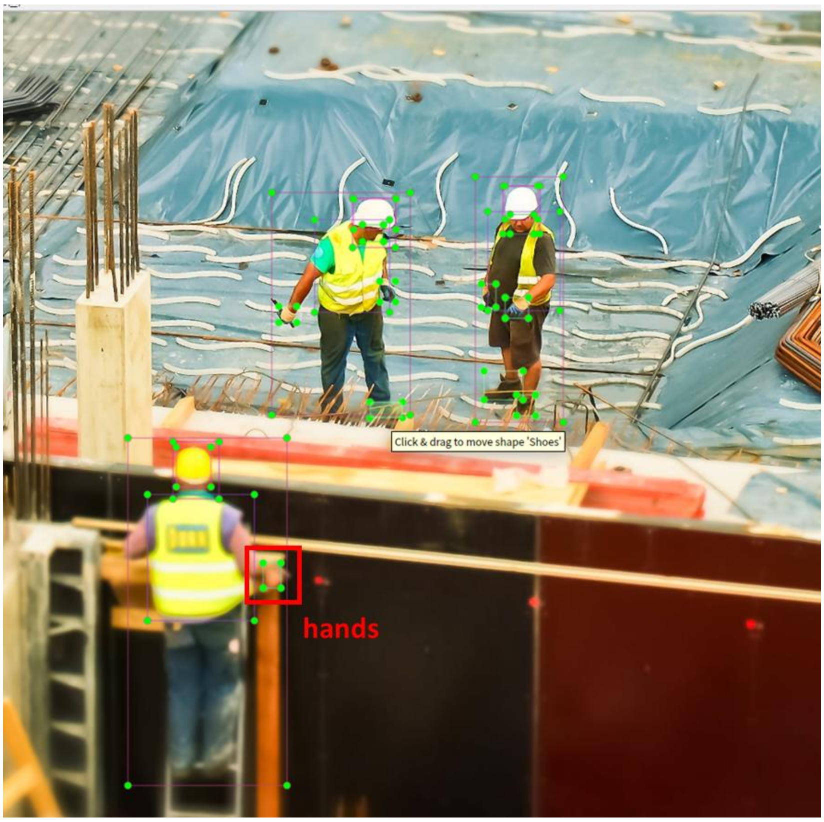


Figure 4-3 Hands annotation in ground truth label.

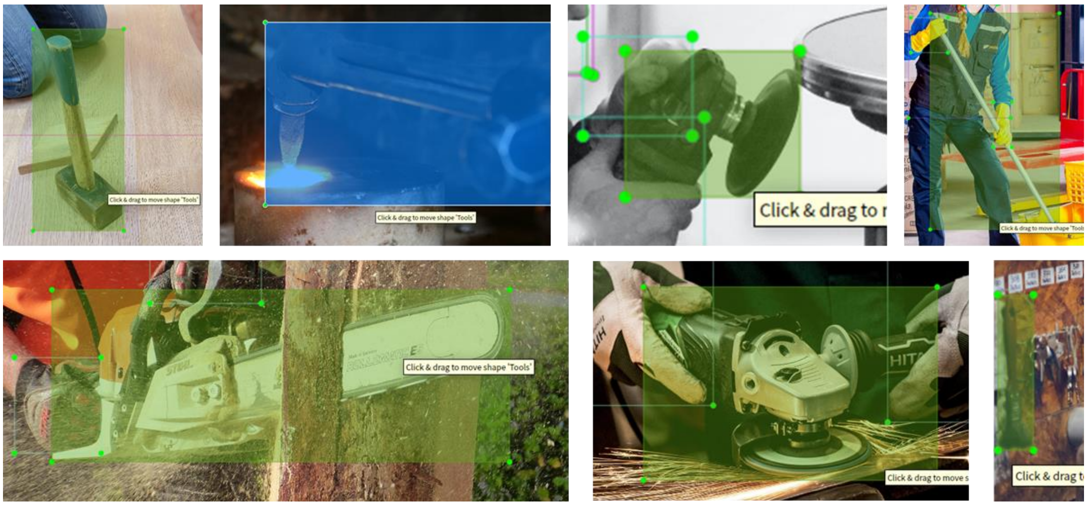

Click & drag to move shape 'Tools'
Click & drag to move shape 'Tools'
Click & drag to r
Click & drag to move shape 'Tools'
Click & drag to move shape 'Tools'
Click & drag to move s
Click & drag t

Figure 4-4 Various types of tools in ground truth labels.

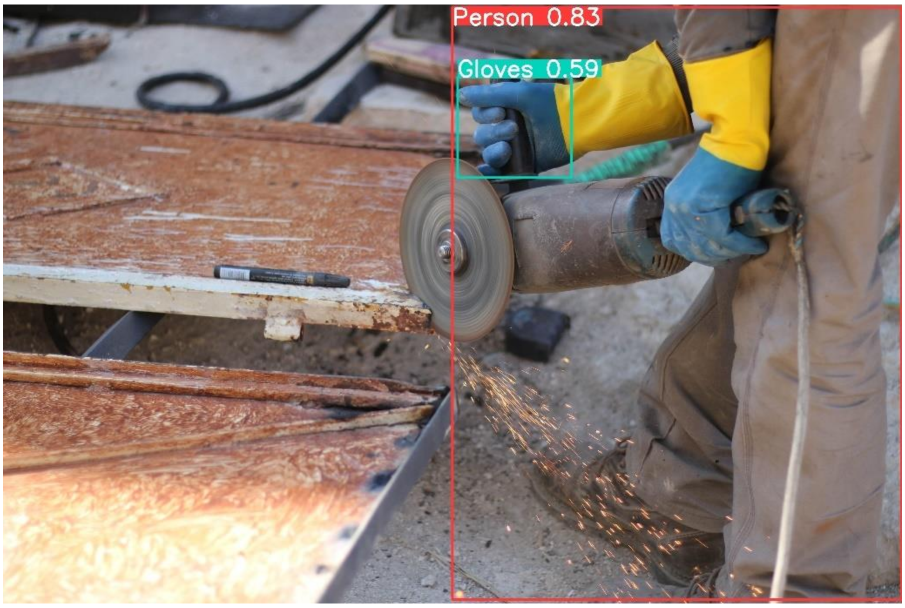


Figure 4-5 Undetected tool.

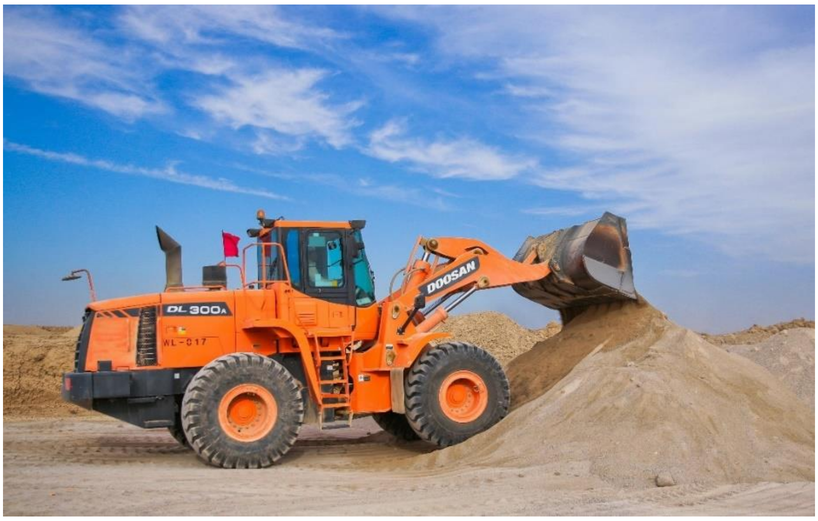


Figure 4-6 Undetected person seated in a device.

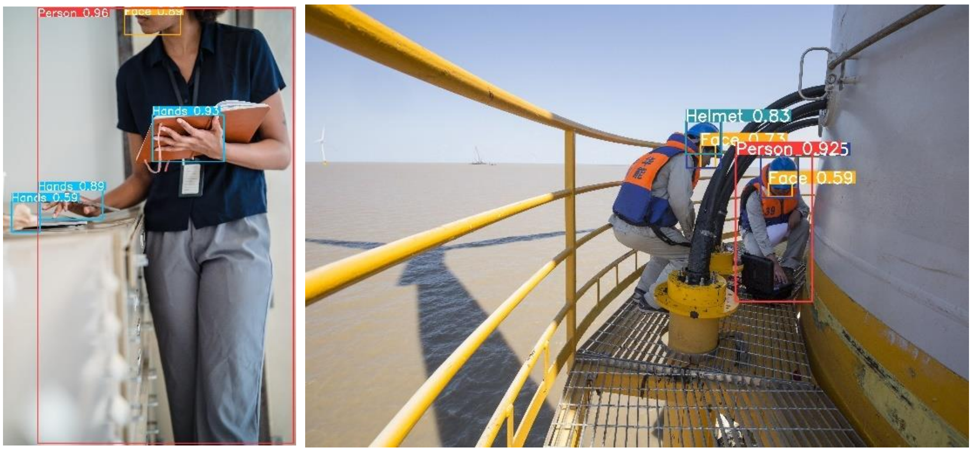


Figure 4-7 Model can accurately identify and detect the face.

5. Reference

Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., and Chen, J. 2024. Detrs beat yolos on real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 16965-16974.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision* (pp. 213-229). Cham: Springer International Publishing.