CORY-Net: Contrastive Res-YOLOv5 Network for Intelligent Safety Monitoring on Power Grid Construction Sites

GUANGYAO PENG1, YIN LEI1, HAOLIN LI1, DI WU1, JUNYUAN WANG2, (Member, IEEE), AND FUQIANG LIU3, (Member, IEEE)
1College of Electronic and Information Engineering, Tongji University, Shanghai, China
2College of Electronic and Information Engineering and the Institute for Advanced Study, Tongji University, Shanghai, China
3College of Electronic and Information Engineering and College of Design and Innovation, Tongji University, Shanghai, China
Corresponding author: Junyuan Wang (e-mail: junyuanwang@tongji.edu.cn) and Fuqiang Liu (e-mail: liufuqiang@tongji.edu.cn)

The work of Junyuan Wang was supported in part by the Shanghai Sailing Program under Grant 20YF1452600, in part by the National Nature Science Foundation of China under Grant 62001330, and in part by the Fundamental Research Funds for the Central Universities.

ABSTRACT In power grid construction projects, ensuring the safety of construction workers by eliminating potential risks has always been important yet difficult. Thanks to the breakthrough in deep learning, it becomes possible to adopt deep learning based object detection technologies to enable intelligent safety monitoring on power grid construction sites. However, due to the complex terrain of the power grid construction site, there is a lack of a dataset of power grid construction scenes. In this paper, a high-quality power grid construction dataset is constructed. In order to reduce the expensive annotation work, we propose to adopt contrastive self-supervised learning to pretrain a feature extraction encoder and integrate it with YOLOv5 by modifying its backbone network for safety-related object detection. The proposed network is referred to as Contrastive Res-YOLOv5 Network (CORY-Net). Experimental results show that contrastive self-supervised pretraining can efficiently improve the object detection performance on the power grid construction dataset and speed up the convergence of the model. With only 1400 pretraining epochs, the proposed CORY-Net outperforms the original YOLOv5 by contributing 1.32% to the mAP@.5 and 3.54% to the mAP@.5:.95. Compared to the supervised object detection benchmarks, the proposed CORY-Net achieves the highest mAP@.5 in 9 out of 22 categories of targets and the highest overall mAP@.5.

INDEX TERMS Contrastive self-supervised pretraining, object detection, power grid construction site, intelligent safety monitoring, YOLOv5

I. INTRODUCTION

Safety monitoring is of paramount importance in power grid construction projects, especially when the working environment is complex and suffers from harsh climate [1], [2]. The traditional way of safety monitoring requires security guards to conduct regular manual inspection and screening, which leads to heavy workload and is of low efficiency. As a result, security risks cannot be completely eliminated on power grid construction sites. Inspired by the appealing performance of deep learning in computer vision tasks, deep learning based object detection technologies could be explored to enable intelligent safety monitoring on power grid construction sites. Specifically, the items related to safety risks, such as heavy equipment, safety vests and cable pulling winches to name a few, can be detected in real time by taking advantage of efficient object detection algorithms, which can be further analyzed to trigger security alarms for instance.

Convolutional neural networks (CNN) based detectors have shown remarkable performance in the field of object detection, which can be divided into three categories: one-stage detectors [3], [4], two-stage detectors [5], [6] and anchor-free one-stage detectors [7]. In contrast to the two-stage detectors based on the region proposal method, the representative one-stage detector, YOLO [4], uses the idea of regression to predict all the categories along with the corresponding confidence and bounding box information, which can speed up the detection greatly yet at the expense of slightly reduced precision. Anchor-free one-stage detectors, such as FCOS [7], were then proposed to avoid the complicated computation regarding the anchor boxes, which...
however were rarely used in multi-objective industrial tasks. Since real-time safety monitoring is highly desirable on the power grid construction sites, the state-of-the-art YOLOv5 appears to be a suitable option thanks to its fast speed.

To apply the YOLOv5 detector to monitor a power grid construction site intelligently, a large dataset of high-quality annotated images captured from diverse practical power grid construction scenes is required. However, due to the complex terrain of power grid construction sites and the harsh climate, it is difficult to collect a massive number of high-quality images covering construction activities comprehensively. Existing datasets were constructed and annotated just for detecting a few specific objects of interest, e.g., helmets [8], safety vests [9] and construction workers [10], which cannot be used to detect other safety-related targets in a complex power grid construction scene. A dataset containing a variety of civil construction scenes was collected in [11], while the power grid construction scenes were not included. Furthermore, manual annotation of images usually requires extensive work [12]. As an image obtained from a power grid construction site usually contains multiple objects in various categories, labeling the objects would be extremely time-consuming and expensive. Due to the difficulties to collect a massive number of high-quality images and the associated time-consuming expensive annotation work, a satisfactory dataset of power grid construction scenes is still missing.

A number of studies devoted to reducing the expensive manual annotation work. A weakly supervised deep learning-based method was proposed in [13] to realize pixel-level cloud detection by training under block-level supervision. The domain-adaptation strategy can fully mine the invariant features between the source domain and the target domain, which can be adopted to solve the image processing tasks when there are no labels. Based on domain adaptation, [14] proposed a new objective function with multiple weakly-supervised constraints to reduce the disadvantageous influence of data shift on the cross-domain remote sensing (RS) image semantic segmentation task. Moreover, the recently proposed contrastive self-supervised learning does not rely on the annotated data and has demonstrated superior performance in computer vision tasks. Many of the contrastive self-supervised learning methods are based on the instance discrimination task [15] that treats each instance as an independent class, and learns the underlying representations from the unlabeled data. Various contrastive self-supervised learning models were proposed, such as MOCO [16] and SimCLR [17], which were shown to be able to achieve comparable performance to the state-of-the-art supervised methods on ImageNet [18]. In addition, by applying contrastive self-supervised learning for pretraining, remarkable transfer performance has been observed in various downstream tasks such as graph classification [19] and object detection [20]. Nevertheless, the transfer capability of contrastive self-supervised pretraining in object detection tasks was mainly evaluated on VOC and COCO [21], [22], while its effectiveness on more complex engineering datasets still remains largely unknown.

Note that, it was revealed in [23] that using datasets in contrastive pretraining similar to the downstream tasks is conducive to representation learning. This motivates us to adopt contrastive self-supervised pretraining to make full use of the unlabeled data collected on power grid construction sites, which could compensate for the lack of a large amount of annotated data for supervised object detection.

In this paper, a dataset of diverse power grid construction scenes is first created. To reduce the data annotation work, contrastive self-supervised learning is adopted to pretrain the feature extraction encoder, based on which YOLOv5 is modified to take advantage of the pretrained encoder and detect safety-related objects on power grid construction sites. The main contributions of this paper are summarized as follows.

- A dataset of 12,000 high-quality images covering diverse practical power grid construction scenes is created, among which 7146 pieces are carefully annotated. The dataset contains 22 categories of objects, including not only the popular objects in the existing datasets such as helmets and safety vests, but also the unique safety-related objects in power grid construction scenes such as cable pulling winch, gin pole and power transmission tower, etc.
- A novel two-stage network, Contrastive Res-YOLOv5 Network (CORY-Net), is proposed, which adopts contrastive self-supervised learning to pretrain a ResNet-50 encoder for feature extraction, and then integrates the resulted ResNet-50 with YOLOv5 to be further trained for object detection on power grid construction sites. Experimental results show that the proposed CORY-Net outperforms the representative fully supervised models on the power grid construction dataset by taking advantage of the extra unannotated data for pretraining. Compared to the state-of-the-art YOLOv5 detector, the proposed CORY-Net contributes 1.32% of the mAP@.5 and 3.54% of the mAP@.5-.95.
- Ablation studies corroborate that the contrastive self-supervised pretraining can efficiently reduce the laborious human-labeling work and the training time. Moreover, more training epochs in the contrastive self-supervised pretraining stage results in better performance of object detection.

II. RELATED WORK

In order to reduce safety risks on construction sites, existing research work mainly focused on detecting the construction workers [10], [24] and whether they are wearing the protective equipment properly [8], [9], [25]–[27]. Faster R-CNN was used to accurately and rapidly detect construction workers in [10]. A dataset including both construction workers and excavators was created in [24], where the iFaster R-CNN algorithm was proposed to automatically detect the construction workers and the heavy equipment that might cause safety
III. OUR METHOD: CORY-NET

In view of the lack of a large dataset of practical power grid construction scenes and the huge cost of data annotation, a Contrastive Res-YOLOv5 network (CORY-Net) is proposed in this paper for safety-related object detection on power grid construction sites.

As depicted in Fig. 1, the proposed CORY-Net consists of two stages: contrastive self-supervised pretraining and supervised object detection. The contrastive self-supervised pretraining is adopted here to make use of the unannotated data to train the encoder, ResNet-50, for efficient feature extraction, which could be then transferred to the downstream object detection task. In the downstream task of supervised object detection, the YOLOv5 model is modified to incorporate the ResNet-50 encoder obtained in the pretraining stage. In the following, let us first introduce the contrastive self-supervised pretraining method adopted in our CORY-Net.

A. CONTRASTIVE SELF-SUPERVISED PRETRAINING

As a representative start-of-the-art contrastive self-supervised learning algorithm, SimCLR is with a simple network structure and can simulate the varying brightness and object size in the images collected on power grid construction sites. Inspired by SimCLR, based on the instance discrimination task, we propose to generate two different augmented versions for every power grid construction image in a batch and train the shared encoder by maximizing the similarity between local features of each image, the proposed model in [30] achieves better transfer performance on ImageNet and COCO. [22] studied the transfer capability of contrastive self-supervised pretraining in two scenarios of few-shot image recognition and facial landmark prediction. Extracting features from different views of the same data in a video was carried out through contrastive self-supervised co-training in [30], which was evaluated in two downstream tasks of behavior recognition and video detection. Despite the above insightful results, whether the contrastive self-supervised pretraining can help with the object detection on power grid construction sites still remains largely unknown.

1) Data Augmentation

The purpose of data augmentation is to alter an original image to produce different augmented versions. Popular data augmentation methods are color transformation and geometric transformation. Color transformation tunes a picture at the pixel level via image noising, grayscale image coloring and color jitter, etc. Geometric transformation simply changes the geometry structure of an image without modifying its pixel level information, including random cropping, flipping and scaling, etc.

As our images are collected on a real power grid construction site, one object usually appears in multiple images.
with different brightness due to weather change and lighting. Moreover, the size of a single object could be very different in different images. To this end, we propose to adopt the composed data augmentation method of random cropping, flipping, scaling and color jitter with some probabilities, which can simulate the interference naturally existed in the power grid construction dataset. We can then expect that the encoder trained in the contrastive self-supervised pretraining stage can deal with the aforementioned interference and thus benefit the downstream object detection task. Fig. 3 shows four images of unfinished towers collected on a real power grid construction site and the corresponding augmented samples generated in the contrastive self-supervised pretraining stage. We can see from the figure that there are obvious differences in the sizes and the brightness of the unfinished towers in the dataset and their augmented versions.

With the proposed data augmentation method, positive pairs, each of which includes two augmented versions of a single image, and a large number of negative pairs of augmented versions of different images in a batch can be produced. Let \( i_m \) denote the \( m \)th sample in the batch \( \{i_m\}_{m=1}^M \), \( t(\cdot) \) and \( t'(\cdot) \) denote two different data augmentation functions. The two augmented versions of sample \( i_m, x_m \) and \( x'_m \), can be written as

\[
x_m = t_m(i_m),
\]

and

\[
x'_m = t'_m(i_m),
\]

respectively, which forms a positive pair.

2) Encoder

The role of encoder in the contrastive self-supervised pretraining is to map an input image into a valid representation of its features. In our CORY-Net, the ResNet-50 network [31] is chosen as the encoder thanks to its ability to balance the network depth and the representation learning capability. A 2048-dimensional feature representation of sample \( x_m, h_m \), can be obtained after encoding as

\[
h_m = \text{ResNet-50}(x_m).
\]

Similarly, the feature representation \( h'_m \) of sample \( x'_m \) is given by

\[
h'_m = \text{ResNet-50}(x'_m).
\]

3) Head

After encoding each image to obtain its feature representation, the head is used to compute the similarity between two augmented images. In our CORY-Net, the nonlinear layers composed of Linear + ReLU + Linear are used to project the 2048-dimensional representation to a 64-dimensional vector as

\[
z_m = \text{Linear}(\text{ReLU}(\text{Linear}(h_m))),
\]

and

\[
z'_m = \text{Linear}(\text{ReLU}(\text{Linear}(h'_m))),
\]

which are then input into the contrastive loss component.
Conv2d at the top of this page is the similarity matrix of all positive negative pair, the network can better distinguish the samples by minimizing the similarity of the representations of every row. Meanwhile, (NT-Xent) is used as the contrastive loss. By maximizing the normalized temperature-scaled cross entropy loss here, the normalized temperature-scaled cross entropy loss

\[ \ell_m = -\log \frac{\exp(s(z_m, z'_m)/\tau)}{\sum_{z_i \neq z_m} \exp(s(z_i, z'_m)/\tau) + \exp(s(z'_m, z_m)/\tau)} \]  

and

\[ \ell'_m = -\log \frac{\exp(s(z'_m, z_m)/\tau)}{\sum_{z_i \neq z_m} \exp(s(z'_m, z_i)/\tau) + \exp(s(z'_m, z_m)/\tau)} \]  

where \( \tau \) denotes a temperature parameter. Finally, the contrastive loss in a batch can be obtained as

\[ loss = \sum_{m=1}^{M} \left( \ell_m + \ell'_m \right) \]  

**B. RES-YOLOv5**

In the second stage of the proposed CORY-Net, a ResYOLOv5 network is designed to fit the characteristics of power grid construction scenes for object detection. In order to take advantage of the encoder obtained in the pretraining stage, as shown in Fig. 4, the ResNet-50 is used as the backbone for feature extraction, which is the same as the encoder in the first stage of contrastive self-supervised pretraining. By doing so, the backbone here can be initialized by the weights obtained in the pretraining stage, which provides prior knowledge of annotated images to the object detection model. The parameters of the backbone can be further fine-tuned by training the supervised learning model.

Since the sizes of the objects in different categories vary greatly in a power grid construction scene, e.g., helmets and power transmission towers, the object detection network needs to output a few feature maps of different sizes to detect both large and small targets. Inspired by YOLOv5, Path Aggregation Network (PANet) is used as the neck and YOLOv3head is chosen as the head to realize predictions across different scales. PANet as the neck between backbone and head is composed of one bottom-up path and one top-down path, which can combine features from various feature layers to deal with the detection of unbalanced sizes. In Fig. 4, three feature maps of sizes 20×20, 40×40, and 80×80 are output from the second, third, and fourth bottleneck blocks with different depths in the ResNet-50 network to connect to the PANet.

The YOLOv3head is used to predict three feature maps of different scales, and predict three boxes for each feature map.
Let $W$ denote the third dimension of the output feature map. If the images are scaled to the resolution of $640 \times 640 \times 3$, the head will output $20 \times 20 \times W$, $40 \times 40 \times W$ and $80 \times 80 \times W$ feature maps. The predicted results include 4 bounding box offsets, 1 objectness prediction, and 22 class predictions. The number of output channels of the feature map is then $W = 3 \times (4 + 1 + 22) = 81$.

### IV. OUR DATASET

A new power grid construction dataset of 12000 images is created, which contains 22 categories of objects, as shown in Fig. 5. Most of the 12000 images are collected on a practical power grid construction site, while the images containing heads are all from the Safety-Helmet-Wearing-Dataset available at [https://github.com/njvisionpower/Safety-Helmet-Wearing-Dataset](https://github.com/njvisionpower/Safety-Helmet-Wearing-Dataset) due to the difficulty to collect the targets of heads without wearing a helmet on a power grid construction site. 7146 out of the 12000 images have been carefully annotated. The training set contains 6125 images and 51609 objects, while the test set contains 1021 images and 7441 objects. This dataset is used in all experiments consistently. Note that these 22 categories of objects are selected as detection targets because they are closely related to potential safety risks during power grid construction. For example, the detection of helmets and safety vests can ensure the safety of the construction workers by checking if the protective equipment is worn properly, while the detection of gin pole and lifting heavy object is essential for the delimitation of dangerous areas on the power grid construction site.

### V. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, experimental results are presented to demonstrate the performance of the proposed CORY-Net on the power grid construction dataset. Ablation studies are performed to evaluate the effectiveness of the contrastive self-supervised pretraining on improving object detection performance in power grid construction scenes. State-of-the-art detectors, YOLOv3, Faster R-CNN, RetinaNet and EfficientDet, are chosen as benchmarks.

#### A. EXPERIMENTAL SETUP

In the contrastive self-supervised pretraining stage, the default hyper-parameters are detailed as follows. The training epochs are fixed at 200, 600, 1000, 1400. The batch size is 256. Adam is used as the optimizer with a learning rate of 0.0003. All experiments are conducted with a 3080Ti GPU. All the 12,000 images in the power grid construction dataset, 

1. Note that diverse images are collected during the whole power grid construction process, which ensures the comprehensiveness of the dataset. As a result, our dataset can also be used for other object detection tasks regarding power grid construction with various aims.
2. Note that the detection performance can be further improved by increasing the batch size in the contrastive self-supervised pretraining stage, since the detection performance benefits more from a larger batch size as revealed in [17].
including both the unlabeled images and the labeled images, are used in the pretraining stage.

In the object detection stage, 100 training epochs are used, the image size is 640 × 640, and the batch size is 32. For the sake of fair comparison, the same setting is applied for all the experiments, which are conducted with a 1080Ti GPU. For our CORY-Net, the backbone ResNet-50 is initialized by the weights obtained in the pretraining stage. The object detection model is then trained and tested on the power grid construction dataset of 7146 annotated images.

As for the scoring criteria, popular metrics including Precision (P), Recall (R), mAP@.5 and mAP@.5:.95 are chosen. mAP@.5 refers to the mAP of intersection over union (IOU) set at threshold 0.5 and mAP@.5:.95 is obtained by averaging the mAP over different thresholds varying from 0.5 to 0.95 with the step size of 0.05.

### B. EXPERIMENTAL RESULTS

1) Performance of CORY-Net:

Table 1 presents the experimental results on our power grid construction dataset. The proposed CORY-Net by integrating contrastive self-supervised pretraining with Res-YOLOv5 for object detection achieves 82.48% Precision, 72.02% Recall, 74.23% mAP@.5 and 43.35% mAP@.5:.95, which outperforms the original YOLOv5 by contributing 1.32% to the mAP@.5 and 3.54% to the mAP@.5:.95. Compared to the Res-YOLOv5 proposed for the downstream object detection, our CORY-Net contributes 1.63% to the Precision, 1.66% to the Recall, 2.47% to the mAP@.5 and 4.05% to the mAP@.5:.95, indicating that the contrastive self-supervised pretraining employed in the CORY-Net can efficiently improve the object detection performance on the power grid construction dataset without any additional cost of human annotation.

Fig. 6 illustrates the detection results in various power grid construction sites.
TABLE 2: Comparison of the proposed CORY-Net with YOLOv3, Faster R-CNN, RetinaNet, EfficientDet and the improved YOLOv5 in [27] on our power grid construction dataset.

| Method          | Epochs | mAP@.5 (all) | mAP@.5 (all) | mAP@.5 :.95 (all) | mAP@.5 :.95 (all) |
|-----------------|--------|--------------|--------------|-------------------|-------------------|
| YOLOv3          | 100    | 63.7         | 75.5         | 36.2              | 72.2              |
|                  |        |              |              |                   |                   |
| Faster R-CNN    | 100    | 69.7         | 91.7         | 38.7              | 72.6              |
|                  |        |              |              |                   |                   |
| EfficientDet    | 16.1   | 98.7         | 69.5         | 0.2               | 0.0               |
|                  |        |              |              |                   |                   |
| RetinaNet       | 58.0   | 72.9         | 89.2         | 39.1              | 78.3              |
|                  |        |              |              |                   |                   |
| Improved YOLOv5 | 70.2   | 73.1         | 85.7         | 36.3              | 79.0              |
|                  |        |              |              |                   |                   |
| CORY-Net        | 100    | 6.1          | 88.8         | 43.1              | 80.7              |

2) Effect of Pretraining Epochs:
Fig. 8 presents the mAP@.5 with the proposed CORY-Net by varying the pretraining epochs. We can see that the mAP@.5 increases as the pretraining epochs increase, indicating that the transfer capability of the contrastive self-supervised pretraining is higher with more training epochs. Moreover, as the mAP@.5 in Fig. 8 has not converged yet, object detection performance could be further improved by increasing the pretraining epochs.

3) Comparison with Benchmarks:
Table 2 presents the mAP@.5 with the proposed CORY-Net, the popular one-stage or two-stage object detection models include YOLOv3, Faster-RCNN, RetinaNet and EfficientDet and the improved YOLOv5 proposed in [27] for helmet detection on construction sites. It can be seen that compared to the start-of-the-art models, our proposed CORY-Net achieves the highest overall mAP@.5, and the best performance for 9 out of 22 categories of objects that need to be detected. Moreover, compared to the improved YOLOv5 network in [27], our proposed CORY-Net achieves higher mAP@.5 for 13 out of 22 categories of objects and contributes 4.0% to the overall mAP@.5. Particularly, significant improvement can be observed in the detection of helmet, ground lead and wireless drum, indicating that the proposed CORY-Net has good detection performance especially for small objects. However, CORY-Net is inferior to other models on the detection of safety vest, gin pole over unfinished tower and anchor rod. We speculate that some characteristics of such targets vary in the power grid construction datasets. For example, anchor rods may be affected by different shooting angles, resulting in differences in their appearance. Safety vests may be partially obscured by the posture of the construction workers and other objects.
VI. CONCLUSION
This paper proposed an object detection model, CORY-Net, by integrating the contrastive self-supervised pretraining and YOLOv5 for intelligent safety monitoring on power grid construction sites. A high-quality dataset of practical power grid construction scenes was also constructed, which includes 22 categories of objects that may pose potential safety risks, which can provide a favorable support for training the proposed CORY-Net. Experimental results showed that our CORY-Net can reduce the amount of required annotated training data and outperforms the supervised model on our power grid construction dataset thanks to the contrastive self-supervised pretraining employed in the CORY-Net. Compared to the YOLOv5 detector, the proposed CORY-Net contributes 1.32% to the mAP@.5 and 3.54% to the mAP@.5:.95. In addition, the proposed CORY-Net achieves higher mAP@.5 for most categories of objects than the state-of-the-art supervised models.

Note that apart from detecting safety-related objects, detection of dangerous areas is also of great importance in intelligent safety monitoring tasks on power grid construction sites, which will be carefully studied in the future by exploring weakly supervised learning methods.

ACKNOWLEDGMENT
The authors would like to thank Mr. Y. Chen for assisting with the data collection on power grid construction sites.

REFERENCES
[1] Z. Yin, Y. Su, G. Luo, Y. Xia, Y. Huang, and Q. Zhang, “Theoretical research based on the application of safety information management in power construction site,” in Proc. IEEE Sustain. Power Energy Conf. (iSPEC), Nov. 2020, pp. 1804–1809.
[2] J. Wu, N. Cai, W. Chen, H. Wang, and G. Wang, “Automatic detection of hardhats worn by construction personnel: A deep learning approach and benchmark dataset,” Autom Construct., vol. 106, no. 102894, Oct. 2019.
[3] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2017, pp. 2980–2988.
[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 779–788.
[5] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. Adv. Neural Inf. Process. Syst., Jun. 2015, pp. 91–99.
[6] Z. Cai and N. Vasconcelos, “Cascade R-CNN: Delving into high quality object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 6154–6162.
[7] Z. Tian, C. Shen, H. Chen, and T. He, “FCOS: Fully convolutional one-stage object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 9626–9635.
[8] L. Huang, Q. Fu, M. He, D. Jiang, and Z. Hao, “Detection algorithm of safety helmet wearing based on deep learning,” Concurreny Comput.: Practice Exp., Feb. 2021.
[9] H. Seong, H. Son, and C. Kim, “A comparative study of machine learning classification for color-based safety vest detection on construction-site images,” KSCE J. Civil Eng., vol. 22, pp. 4254–4262, Sep. 2018.
[10] H. Son, H. Choi, H. Seong, and C. Kim, “Detection of construction workers under varying poses and changing background in image sequences via very deep residual networks,” Autom Construct., vol. 99, pp. 27–38, Mar. 2019.
[11] A. Xue-hui, Z. Li, L. Zuguang, W. Chengzhi, L. Pengfei, and L. Zhiwei, “Dataset and benchmark for detecting moving objects in construction sites,” Autom Construct., vol. 122, no. 103482, Feb. 2021.
[12] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, “A survey on contrastive self-supervised learning,” Technologies, vol. 9, no. 1, p. 2, Oct. 2021.
[13] Y. Li, W. Chen, Y. Zhang, C. Tao, R. Xiao, and Y. Tan, “Accurate cloud detection in high-resolution remote sensing imagery by weakly supervised deep learning,” Remote Sensing of Environment, vol. 250, no. 112045, Dec. 2020.
[14] Y. Li, T. Shi, Y. Zhang, W. Chen, Z. Wang, and H. Li, “Learning deep semantic segmentation network under multiple weakly-supervised constraints for cross-domain remote sensing image semantic segmentation,” ISPRS Journal of Photogrammetry and Remote Sensing, vol. 175, pp. 20–33, May 2021.
[15] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, “Unsupervised feature learning via non-parametric instance discrimination,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 3733–3742.
[16] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2020, pp. 9792–9798.
[17] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in Proc. Int. Conf. Mach. Learn., Feb. 2020, pp. 1597–1607.
[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 248–255.
[19] J. Zeng and P. Xie, “Contrastive self-supervised learning for graph classification,” AAAI, vol. 35, no. 12, pp. 10824–10832, May 2021.
[20] S. Chen, X. Nie, D. Fan, D. Zhang, V. Bhat, and R. Hamid, “Shot contrastive self-supervised learning for scene boundary detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Apr. 2021, pp. 9796–9805.
[21] A. Islam, C.-F. Chen, R. Panda, L. Karlinsky, R. Radke, and R. Feris, “A broad study on the transfer ability of visual representations with contrastive learning,” arXiv e-prints, p. arXiv:2103.13517, Mar. 2021.
[22] Q. Zhao, Z. Wu, R. W. H. Lau, and S. Lin, “What makes instance discrimination good for transfer learning?” in Proc. Int. Conf. Learn. Represent. (ICLR), Jun. 2021.
[23] K. Kotar, G. Ilharco, L. Schmidt, K. Ehsani, and R. Mottaghi, “Contrasting contrastive self-supervised representation learning models,” arXiv e-prints, p. arXiv:2103.14005, May 2021. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2021arXiv210314005K.
[24] W. Fang, L. Ding, B. Zhong, P. Love, and H. Luo, “Automated detection of workers and heavy equipment on construction sites: A convolutional neural network approach,” Adv. Eng. Informatics, vol. 37, pp. 139–149, Aug. 2018.
[25] M. Saud, A. Hakim, A. Ahmad, A. Shaker, M. Hanafi, A. Narzullaev, and M. Ifwat, “Image detection model for construction worker safety conditions using Faster R-CNN,” Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 9, Nov. 2020.
[26] Y. Zhao, Q. Chen, W. Cao, J. Yang, J. Xiong, and G. Gui, “Deep learning for risk detection and trajectory tracking at construction sites,” IEEE Access, vol. 7, pp. 30905–30912, Mar. 2019.
[27] S. Tan, G. Lu, Z. Jiang, and L. Huang, “Improved yolov5 network model and application in safety helmet detection,” in Proc. IEEE Int. Conf. Intell. Safety Robotics (ISR), Dec. 2020, pp. 330–333.
[28] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, “Learning deep representations by mutual information estimation and maximization,” in Proc. Int. Conf. Learn. Represent. (ICLR), Aug. 2019, pp. 2–25.
[29] O. J. Hénaff, S. Koppula, J.-B. Alayrac, A. van den Oord, O. Vinyals, and J. Carreira, “Efficient visual pretraining with contrastive detection,” arXiv e-prints, p. arXiv:2103.10957, Mar. 2021. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2021arXiv210310957H.
[30] T. Han, W. Xie, and A. Zisserman, “Self-supervised co-training for video representation learning,” in Proc. Int. Conf. Neural Inf. Process. Syst., vol. 33, Oct. 2020, pp. 5679–5690.
[31] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun 2016, pp. 770–777.