Development of a Petrochemical Plant Pipe Counting Method Based on Improved Mask R-CNN and Transformer

Rong-Lu Hong[1], Jin-Bin Im[1], Sang-Jun Park[1], En-Lian Zhang[1], Nomgon Ochirsuren[1], Young-Gun Beak[1], Shin-Hyun Kang[1], Seong-Won Chung[1], Cheol-Su Lee[2], Seunghyeon Wang[1] and Ju-Hyung Kim[1]

[1] Department of Architectural Engineering, Hanyang University, Seoul, Republic of Korea
[2] UND Co., Ltd., Suseong-gu, Daegu, Republic of Korea
abraham0814@hanyang.ac.kr, jbim0202@hanyang.ac.kr, sangjpark@hanyang.ac.kr, eunryeon@hanyang.ac.kr, nomgon.ochirsuren53@gmail.com, byggun222@hanyang.ac.kr, shinhuny@naver.com, jsw7019@naver.com, bruce@unde.co.kr, wsh1019@hanyang.ac.kr , kcr97jhk@hanyang.ac.kr

Abstract

In petrochemical plant construction, pipe installation critically the affects overall progress and costs, making accurate on-site pipe quantification essential. However, the wide range of pipe diameters, high-density layouts, and complex site conditions poses significant challenges for automated counting methods. To address these issues, this study constructs the "PetroPipe" dataset, comprising 2,380 images (340 originals and 2,040 augmented) with pipes ranging from 0.75 to 60 inches in diameter under various resolutions. Multiple data augmentation strategies and a multi-stage training process enhance robustness under dynamic construction conditions. Building on this dataset, a pipe counting method is proposed that integrates instance segmentation with a counting head and a transformer encoder. Based on the Mask R-CNN framework, the model applies a discrete-bin classification approach to achieve accurate global quantity predictions, while maintaining high-quality instance-level detection and segmentation. The experimental results demonstrate that the developed model outperforms the standard Mask R-CNN model on the test set, increasing the counting accuracy from 65.55% without a counting head to 85.50%. When evaluated on an unlabelled 18-second video captured in a real-world scenario, it attained an accuracy of approximately 98.1%, demonstrating its capacity to adapt to practical industrial conditions. These results indicate that the proposed approach effectively accommodates varying pipe diameters, dense layouts, and complex backgrounds. Consequently, it provides a scalable, efficient pipe counting, progress monitoring, and resource optimization in petrochemical construction environments.

Keywords –

Petrochemical plant pipe; Instance Segmentation; Counting; Transformer

1 Introduction

During construction of petrochemical plants, pipe installation is a crucial element of the overall construction schedule. It is a prominent component of workload and cost distribution [1]. Statistical evidence suggests that pipe installation often dictates the overall project timeline, with materials and labor accounting for approximately 71% of the total pipeline construction costs [2]. Given the large volumes and varied specifications of piping materials, this key stage significantly influences project outcomes. To ensure smooth progress, project managers must accurately and efficiently record and verify the on-site quantities, types, specifications, and manufacturers of these materials [3].

However, the widely used material management approaches, including barcodes, radio frequency identification (RFID) tags, and manual counting, frequently face significant challenges in practice. On the one hand, manual verification and data entry are time-consuming and labour-intensive, and they are easily affected by human errors and environmental interferences. In addition, RFID and barcode systems require special hardware and pre-applied labels, conditions that cannot always be satisfiedin complex and rapidly changing construction environments [4]. Real-time tracking and frequent updates of material information are often required. Under these conditions, traditional methods fail to provide the speed and flexibility required for modern detail-oriented project management [3].

In this context, rapid advances in computer vision and image processing technologies have provided new opportunities for digitalization and intelligent management of construction sites. Previous studies have successfully applied deep learning-based visual analysis methods to tasks such as crack detection, structural damage assessment, and construction quality control [5,

6], indicating that using image data for automated construction material counting is both feasible and competitive. Compared with conventional labelling methods, image processing requires no additional hardware markers and can extract information in real time from video streams, thereby significantly improving operational efficiency and data accuracy [5]. Nevertheless, the direct application of image processing techniques to pipe counting in petrochemical plants remains challenging. Piping materials are frequently stacked in high densities, differ in size and shape, and are often partially obscured [3]. Consequently, traditional detection methods that rely solely on bounding boxes struggle to delineate pipe boundaries accurately in dense scenes, leading to counting errors [7]. Moreover, the complexity of the environment and variability in data distributions highlight the urgent need to enhance model robustness and generalization.

To address these challenges, instance segmentation provides a more direct technical pathway for improving counting accuracy and detection reliability. Among these, Mask R-CNN, as a classic and widely used deep-learning method, has been extensively explored and validated in counting tasks for plants, vehicles, and animals. However, it exhibits insufficient accuracy and limited generalization when handling complex scenarios characterized by high target density, significant scale variations, and severe occlusions [8–10]. By contrast, transformers have demonstrated initial success in capturing global context, integrating multi-scale information, and enhancing feature representations in other counting tasks [11, 12]. Incorporating a transformer into the instance segmentation and counting framework may improve the adaptability and generalization of the model in petrochemical plant settings characterized by diverse pipe diameters, high densities, and significant scale variations. This study addresses the current gaps in the literature related to these aspects.

This study extends the classical Mask R-CNN instance segmentation model by integrating a Transformer-based counting head into the existing framework. This integration enables precise multi-object instance segmentation while directly estimating the overall number of targets. Using a multi-task joint training strategy, the model achieves coordinated optimization of detection, segmentation, and counting at the feature level, thereby improving its robustness and generalization in complex construction scenarios. Moreover, the proposed approach adapted for other types of densely distributed construction materials and can be seamlessly integrated into on-site digital information management systems, ultimately providing more efficient, accurate, and intelligent support for project scheduling, resource allocation, and cost optimization.

2 Related Work

Over the past decade, traditional computer vision techniques based on image processing have progressed in counting construction materials. Thammasorn et al. (2013) proposed a template matching method for counting rectangular and circular steel pipes, achieving high accuracy when the camera angle is relatively fixed and the main targets occupy most of the image [13]. Ghazali et al. (2017) combined the Hough transform with post-processing and morphological operations to detect and count circular and rectangular steel materials, thereby demonstrating initial practical outcomes [14]. Liu et al. (2019) employed the Prewitt operator to extract oriented gradient features and used support vector machine (SVM) to locate and count bundled rebars, achieving an average accuracy of approximately91% in their experiments [15].

However, traditional image processing and feature engineering methods are sensitive to environmental conditions. Factors such as the camera angle, illumination changes, steel material size, cross-sectional shape, and background clutter, can cause significant performance fluctuations [3]. In actual construction sites, uneven lighting, weather variations, and increasingly complex background textures further compromise the stability and generalizability of such methods [5]. This highlights the difficulty of designing sufficiently robust features that can handle the diversity and uncertainty of on-site environments using traditional approaches.

In recent years, deep learning methods, including convolutional neural networks (CNNs), have gained wide acceptance and validation in the visual analysis of construction engineering [16]. Compared with traditional image processing techniques, these approaches can handle complex scenes more effectively while improving both accuracy and robustness, thereby offering strong technical support for construction informatization and intelligent decision-making [5, 17]. Motivated by these advances, researchers have started exploring deep learning techniques for automated counting of construction materials. For example, Fan et al. (2019) reframed the task of counting rebars as an image classification problem by classifying local patches within an image [17]. However, this approach requires the examination of each patch individually, which prolongs the inference time. In addition, its reliance on parameters such as target radius restricts its applicability in dynamic and ever-changing construction environments. In efforts to improve practical performance, Wang et al. (2023) integrated UAV imagery with Faster R-CNN and trained 384 model variants on both original and augmented datasets collected under real construction conditions [5]. Their best model achieved an accuracy of 94.16% at AP50. Similarly, Li et al. (2022) developed a precise counting model using YOLOv3 and released a large-

scale steel pipe dataset covering diverse on-site conditions, thereby providing abundant resources for research [3].

To achieve deeper feature representation and stronger global context capture, Bai et al. (2023) introduced the CounTr model, which uses a pure Transformer architecture with multi-head self-attention. By using a hierarchical self-attention decoder to fuse multi-scale and global information, CounTr delivered significantly better performance than traditional CNN-based approaches on both the UCF_CC_50 high-density crowd dataset and the FDST diversified dataset [11]. Feng et al. (2023) presented a pig counting and localization method based on density maps. Their approach employs a lightweight hybrid network that integrates selective convolution and a Vision Transformer to generate density maps and then uses an improved K-means algorithm to locate individual targets on these maps. The experimental results showed a low counting error (MAE=0.726), along with precision and recall rates of 88.22% and 86.02%, respectively, making it suitable for large-scale and crowded scenarios [12]. However, in the context of pipe counting, the current research remains largely confined to object detection and classification, with insufficient in-depth exploration of the integration of instance segmentation and Transformer-based techniques. Instance segmentation provides independent, pixel-level masks and precise boundary information for each pipe, thereby effectively enhancing the counting accuracy and stability in complex backgrounds [7]. It also provides a solid foundation for subsequent pipe layout planning and optimization. Incorporating Transformers into this framework is expected to strengthen the capture of the global context and integration of multi-scale features. These improvements facilitate better adaptation and generalization to the challenging conditions present in petrochemical plants, including a wide range of pipe diameters, high densities, and significant scale variations.

In summary, applying a counting method based on instance segmentation and Transformers for pipe counting in petrochemical plants is both a logical and potentially valuable. This approach not only compensates for the shortcomings of existing research under complex industrial conditions, but can also provide more comprehensive and robust technical support for achieving more accurate counting and spatial layout optimization in future industrial applications.

3 Material and Methods

3.1 Data collection

The effective training of deep learning models often relies on large and diverse image datasets. However, research on counting construction materials, particularly pipes, has been limited by the lack of large-scale data resources.

To address this shortcoming, this study constructed a comprehensive dataset designed for a petrochemical plant environment, referred to as the “PetroPipe” dataset. This dataset is publicly available on Kaggle and comprises 340 images with 14,855 pipe instances [18]. All images were captured with an iPhone 14 Pro under various conditions and from multiple angles, increasing the data diversity and improving the adaptability of the model. As shown in Figure 1, the pipes cover diameters ranging from 0.75 60 inches, ensuring that the trained models will be more robust in real-world scenarios. The dataset includes images of three different resolutions (3061×3024, 3024×4032, and 4032×3024), reflecting the variety of image sizes encountered in industrial environments. The inclusion of multiple resolutions helps the model maintain strong generalization performance under varying imaging conditions.

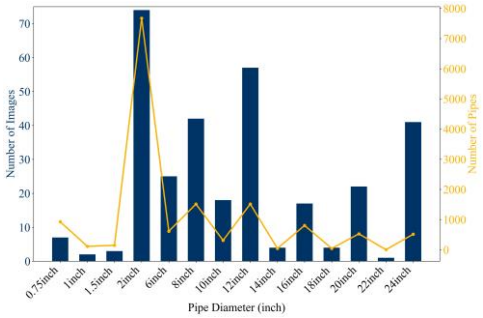

Figure 1. Distribution in the PetroPipe dataset

Figure 2. Example of polygon-based annotations

For data annotation, polygon-based annotations were created for each pipe instance using the Labelme tool. These annotations were then converted into the COCO format to ensure direct compatibility with the Mask R-CNN model [19]. Figure 2 shows an example of an annotated image. The performance of the model was evaluated under near-real-time monitoring conditions using an 18-second video containing multiple pipe instances

3.2 Data augmentation

Vrious data augmentation operations were applied to each original image to enhance the robustness of the model in complex real-world scenarios. As shown in Figure 3, these include blur, brightness, color, contrast, horizontal flip, and sharpness adjustments. Such strategies effectively increase the diversity of the dataset, enabling the model to adapt better to different lighting conditions, camera angles, and image quality variations [20]. A total of, 2,040 augmented images were generated, and together with the original 340 images, yielded a dataset of 2,380 images. The dataset was then split into training, validation, and test sets in a ratio of 6:2:2 to ensure reliable training and evaluation.

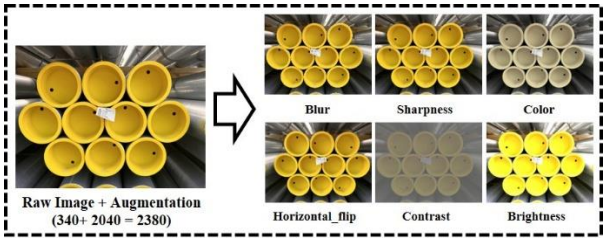

Figure 3. Examples of the data augmentation operations

3.3 Model Architecture and Counting Head Design

The model architecture and experimental workflow are shown in Figure 5. To achieve rapid and accurate global object counting, this study integrates a counting head into the standard Mask R-CNN framework. This integration enables the model to produce high-quality instance-level segmentation while estimating the total number of targets in the entire image from a global perspective. The base structure uses a ResNet-50 architecture combined with a feature pyramid network (FPN) as the backbone. And a region proposal network (RPN) generates initial proposals. Each proposal undergoes region of interest(ROI) align and box head feature extraction, resulting in a 1024dimensional feature representations. By leveraging these representations, the box regression and classification branches provide precise object detection and localization, whereas the mask branch delivers instance-level semantic segmentation [8]. However, the conventional Mask R-CNN architecture does not directly count the total number of objects across an entire image.

As shown in Figures 4 and 5, the proposed counting head is introduced at the output of the box head to address this limitation. Instead of simply averaging the ROI features, a global token is extracted from a lower-level FPN feature map (P2 layer) using global average pooling. The global token is concatenated with all the ROI features (ROI Tokens) to form a sequence comprising one global token and multiple ROI tokens. The sequence is then fed into a Transformer encoder, where multi-head self-attention and feed-forward layers achieve deep fusion of global and instance-level features. In the Transformer output's, the first token corresponds to the global token, and can be regarded as a fused global feature representation.

A linear classifier (count_classifier) subsequently maps the fused global features to scores over a set of predefined discrete intervals representing plausible target counts (e.g. dividing the range of 0–50 into several bins). Instead of predicting a continuous number directly, the model selects the bin with the highest score to infer the total object count. This classification-based strategy stabilizes the training process and enhances interpretability. Moreover, the counting head imposes a minimal additional computational overhead, offering an efficient and innovative solution that integrates both local and global features for object detection and counting tasks.

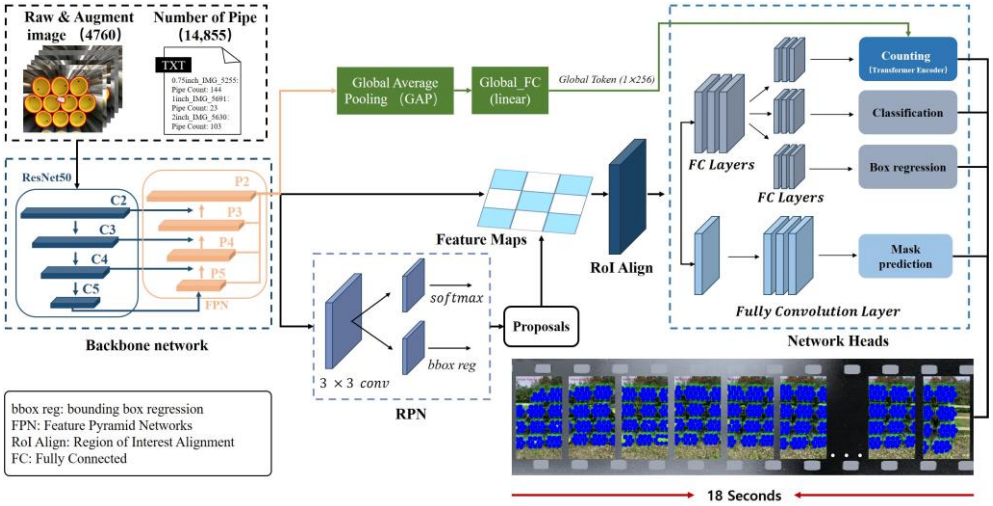

Figure 5. Model architecture and experimental workflow

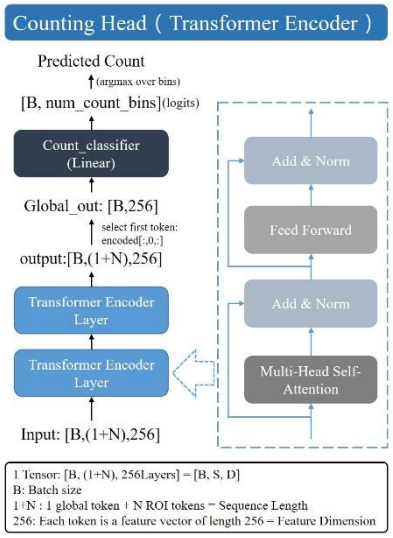

Figure 4. Internal structure of the counting head

3.4 Experimental Settings

This study employed a multi-stage multi-task optimization strategy when integrating the counting head, ensuring steady improvements in the detection and segmentation performance. Initially, the backbone weights were obtained from a ResNet-50+FPN model pre-trained on the COCO dataset. To maintain stable convergence and avoid abrupt parameter shifts, the backbone parameters were frozen during the early training stage (the first two epochs), and only the ROI heads and counting head parameters were updated. At this point, the loss weight for the counting branch was set relatively low (0.1) to prevent the counting task from dominating the training process too soon.

Once the detection, segmentation, and counting tasks achieved a preliminary balance, as indicated by completion of the freezing phase, the backbone parameters were unfrozen, allowing the entire network to be fine-tuned. This step enabled the model to adapt better to the characteristics of the target data and enhance its overall performance. Thus, to overcome the two major challenges in multitask training, namely, the interference of the counting branch with the overall network and the imbalance of loss weights, we ensured in the early stage, through a lower counting loss weight and dynamically adjusted learning rates, that the counting head does not cause excessive gradient impact when jointly training with the detection and segmentation tasks. In the later stage, when the backbone was unfrozen and fine-tuned jointly, the counting loss weight was moderately increased and a staged learning rate schedule was adopted, enabling the counting head to acquire richer semantic information without undermining the previously learned detection and segmentation capabilities. During training, the AdamW optimizer and a StepLR scheduler were employed to gradually reduce the learning rate in phases, ensuring a smooth transition from rapid initial convergence to more refined adjustments. Owing to this multi-stage approach, the model preserved the detection and segmentation improvements after incorporating the counting head. In addition, it achieved a significant increase in the global counting precision, demonstrating effective multi-task optimization. Details of the experimental environment and parameter settings are listed in Table 1.

Table 1. Details of the experimental environment and parameter settings

Parameter	Description
Operating System	Windows 11 Pro
Hardware	GPU: RTX 3060 (12GB), CPU: AMD Ryzen 5 5600 RAM: 32GB
Development Environment	PyTorch 2.5.1, Python 3.9, CUDA 12.4
Optimizer	AdamW (lr=1e-5, weight_decay=1e-4)
LR Scheduler	StepLR: halving lr every 5 epochs; total 100 epochs
Initial Weights	MaskRCNN_ResNet50_FPN_ Weight. COCO_V1
Training Strategy	Batch=2; First 2 epochs: freeze backbone (train ROI & counting head)

3.5 Evaluation Metrics

To objectively evaluate the performance of the instance segmentation model proposed in this paper, we define the evaluation metrics at two scales: bounding box and mask. Specifically, the metrics average precision (AP) and mean intersection over union (mIoU) were used to assess the model's overall detection ability and its fine localization performance. We calculated $\boldsymbol{AP_{50}}$ and $\boldsymbol{AP_{75}}$ at IoU thresholds of 0.5 and 0.75, respectively. The AP metric is defined as

$$\boldsymbol{AP} = \int_0^1 \boldsymbol{p}(\boldsymbol{r})\boldsymbol{dr} \quad (1)$$

where $\boldsymbol{p}(\boldsymbol{r})$ is the interpolated precision value under the precision–recall curve. The AP computed at the bounding box scale is used to measure the overall localization ability of the model, whereas the AP at the mask scale further reflects the fine segmentation performance of the model at pixel level [5,6]. The AP metric has been widely applied in the field of instance segmentation because it effectively evaluates the overall detection and fine localization capabilities of the model,

and intuitively and comprehensively reflects the recognition performance of the model under varying degrees of strictness [6,8]. In addition, we computed mIoU at both the bounding box and mask scales using the general formula defined as follows:

$$\boldsymbol{mIoU} = \frac{1}{N}\sum_{i=1}^{N}\frac{R_{pred}^{i} \cap R_{gt}^{i}}{R_{pred}^{i} \cup R_{gt}^{i}} \quad (2)$$

where $\boldsymbol{R_{pred}^{i}}$ and $\boldsymbol{R_{gt}^{i}}$ are the predicted and ground-truth regions for the ith target, respectively. This formula is applicable to both the bounding box and mask scales. At the bounding box scale, the mIoU emphasizes the overall localization ability of the model, whereas at the mask scale, it more precisely reflects the model's ability to delineate target boundaries at pixel level [5,8]. Therefore, computing the mIoU at both scales can comprehensively demonstrate the performance advantages of the model at different granularity levels. To evaluate the object counting performance, we compared the number of instances predicted by the model with the counting module to those predicted by the native detection output without parameter adjustments to assess the contribution of the counting module to counting accuracy and its impact on overall model performance. We designed a multitask joint loss function that integrates detection (classification and bounding box regression), mask segmentation, and object counting.

$$\boldsymbol{L_{Total}} = \boldsymbol{L_{cls.}} + \boldsymbol{L_{bbox\,reg}} + \boldsymbol{L_{mask}} + \boldsymbol{\lambda L_{count}} \quad (3)$$

where λ is used to control the influence of the counting loss term, thereby achieving collaborative optimization among the different task losses. Compared with other potential metrics, such as the Dice coefficient or pixel accuracy, the AP and mIoU metrics offer more comprehensive, intuitive, and stable advantages, enabling an effective and detailed evaluation of the precision and generalization performance of the instance segmentation model.

4 Results

4.1 Results of Training and Validation

During model training, both the training and validation losses were monitored, as shown in Figure 6. Although the initial plan was to train for 100 epochs, the model adapted quickly. Initially, the training loss was 1.11, and the validation loss was 0.71. By the 5th epoch, the training loss decreased to 0.23, and the validation loss was 0.22, indicating effective feature learning. After the 15th epoch, both training and validation losses stabilized at 0.14. However, given that the transformer model is strongly dependent on global information and positional features, it generally requires a longer training period to fully integrate the features and achieve stable convergence. Therefore, in this experiment, the training process was extended to 100 epochs to ensure that the model achieved stable, low-loss convergence and exhibited strong generalization capabilities under the chosen optimization strategy.

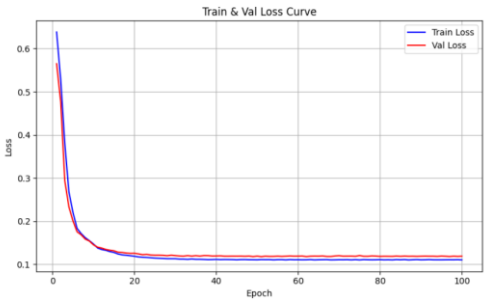

Figure 6. Training and validation losses

4.2 Model Evaluationusing Test Images

Comparative experiments were conducted to evaluate the performance of the standard Mask R-CNN (ResNet-50) and the proposed model under identical datasets and configuration settings. As shown in Table 2, the test results indicate that our model slightly outperformed the standard Mask R-CNN when evaluated at both the bounding box and mask scales. Although the improvements are modest, these results demonstrate that our model exhibits greater stability in terms of capturing object details and segmentation accuracy.

Table 2. Comparison of model performance

Model	Mask R-CNN		Our Model	
Label	BBox	Mask	BBox	Mask
AP 50	97.71%	97.26%	98.07%	97.45%
AP 75	97.50%	96.98%	97.89%	97.20%
mIOU	85.56%	76.47%	87.32%	76.54%
Counting Accuracy	65.55%		85.50%	

For the counting task, Mask R-CNN employs a shared confidence scoring mechanism for detection boxes and segmentation masks; thus, the counting accuracy is computed as a unified metric rather than separately for different scales. Furthermore, no confidence threshold was set in this study to ensure that all predicted instances were counted. The experimental results showed that the counting accuracy of the standard model was 65.55%. By contrast, Our model incorporates an independent counting module, which integrates a transformer encoder to consolidate global information and instance features, and employs a discrete bin classification strategy to directly output the instance count. Owing to this improvement, the counting accuracy of the new model increased to 85.50%, indicating that the introduced counting head provides more stable global statistical

information for the model, thereby overcoming the potential instability of relying solely on detection outputs.

However, as shown in Figures 7(c) and 7(d), the segmentation and counting accuracies of our model remained limited in scenarios with extremely dense and small-diameter pipes. Such conditions involve blurred boundaries, severe overlap, and low feature distinctiveness, thereby increasing the difficulty of detection and counting. By contrast, as shown in Figures 7(a) and 7(b), our model achieved nearly 100% counting accuracy in scenarios with fewer pipes. These findings indicate that although the proposed approach is effective in less congested environments, further improvements are required to address highly crowded and complex conditions.

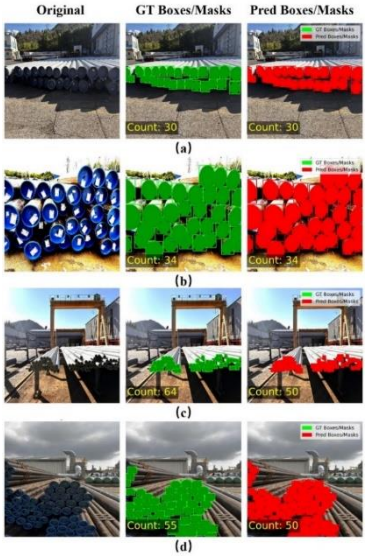

Figure 7. Our model's pipe segmentation and counting examples under different densities

4.3 Video Testing

To further assess the practicality and adaptability of the improved model, inference was conducted on an unlabeled, real-world video (608×1080, 18s, 30 FPS) containing 460 pipes characterized by a high density and small object size. Operating in the evaluation mode with float32 precision, the model processed each frame without additional training. An IoU-based matching strategy was employed to avoid multiple counts of the same pipe, either by updating existing Identifications (IDs) or assigning new IDs as required.

By applying frame-by-frame inference, the model detected 469 pipes, achieving approximately 98.1% counting accuracy compared with the actual total of 460. This result demonstrates that even without annotated training data, the counting head approach can significantly enhance the counting precision in real-world scenarios. Although the average per-frame inference time was about 0.4176 seconds (approximately 2.39 FPS) and thus did not meet strict real-time requirements, the method maintained a relatively high degree of accuracy under dense conditions.

The visualization results (Figure 8(a) show the initial frame and Figure 8(d) shows the final frame) indicate stable detection and counting performance, even in complex, tightly arranged scenes. These findings provide a valuable reference for future research and practical applications of unannotated video monitoring.

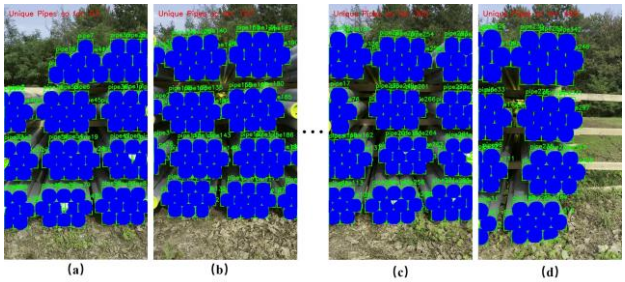

Figure 8. Sample of sequential frames

5 Conclusions

This study focused on pipe counting issues in petrochemical plants. It proposes a counting method that integrates instance segmentation with a Transformer, and through a multi-task joint training strategy, achieves outstanding performance in complex, densely populated, and highly variable construction site environments. The experimental results showed that the model achieving an AP50 of 98.07% (BBox) / 97.45% (mask), an AP75 of 97.89% (BBox) / 97.20% (mask), and an mIoU of 87.32% (BBox) / 76.54% (mask), and inceased the counting accuracy from 65.50% (without the counting head) to 85.50%. When applied to real-world, unlabeled video data, the model demonstrates strong generalization capability and achieves a counting accuracy of about 98.1%, confirming its practical value in real scenarios. Nevertheless, certain limitations remain in practical applications. Although the proposed model demonstrated robust generalization ability and counting accuracy in practical application scenarios, there is still room for improvement in the detection performance of densely packed, small-diameter targets and inference speed. Future research, in addition to further expanding the dataset scale and diversity, should consider optimizing the internal structure of transformers by incorporating multiscale attention mechanisms or local-global feature fusion strategies to enhance the recognition of dense and small-sized targets. The current backbone network can be replaced with a more efficient and lightweight CNN or transformer architectures (e.g., EfficientNet, Swin-Tiny, and MobileViT) and integrated with model compression and pruning techniques to further reduce computational

costs and inference time while preserving accuracy, thus better addressing the practical deployment requirements in industrial settings. This study provides a technological foundation for pipe counting and information management in petrochemical facilities, establishing a solid basis for further improvements in speed, data expansion, and framework optimization, thereby contributing to the intelligent and automated advancement of this field.

Acknowledgement

This research was funded by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (Ministry of Science and ICT) (grant number RS202400356697).

References

[1] G. Qin, H. Ding, X. Liu, H. Zhang, and G. Ma, "Construction management of petrochemical process pipeline installation," *Smart Construction Research*, 2018

[2] Z. Rui, P. A. Metz, D. B. Reynolds, G. Chen, and X. Zhou. Historical pipeline construction cost analysis. *International Journal of Oil, Gas and Coal Technology*, 4(3):244–263, 2011.

[3] Y. Li and J. Chen. Computer vision–based counting model for dense steel pipe on construction sites. *Journal of Construction Engineering and Management*, 148(1):04021178, 2022.

[4] K. K. Duan and S. Y. Cao. Emerging RFID technology in structural engineering–A review. *In Structures*, 28: 2404–2414, 2020.

[5] S. Wang, M. Kim, H. Hae, M. Cao, and J. Kim, The Development of a Rebar-Counting Model for Reinforced Concrete Columns: Using an Unmanned Aerial Vehicle and Deep-Learning Approach, *Journal of Construction Engineering and Management*, 149(11), 2023.

[6] Z. Liu, J. K. Yeoh, X. Gu, Q. Dong, Y. Chen, W. Wu, and D. Wang. Automatic pixel-level detection of vertical cracks in asphalt pavement based on GPR investigation and improved mask R-CNN. *Automation in Construction*, 146:104689, 2023.

[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. *In Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, Venice, Italy, 2017.

[8] Y. Shao, X. Guan, G. Xuan, H. Liu, X. Li, F. Gu, Z. Hu, Detection of Straw Coverage under Conservation Tillage Based on an Improved Mask Regional Convolutional Neural Network (Mask R-CNN). *Agronomy*, 14:1409, 2024.

[9] B. Xu, W. Wang, G. Falzon, P. Kwan, L. Guo, G. Chen, A. Tait, D. Schneider, Automated cattle counting using Mask R-CNN in quadcopter vision system, *Computers and Electronics in Agriculture*, 171:105300, 2020.

[10] H. Tahir, M. Shahbaz Khan, M. Owais Tariq, Performance Analysis and Comparison of Faster R-CNN, Mask R-CNN and ResNet50 for the Detection and Counting of Vehicles, *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pages 587–594, Greater Noida, India, 2021.

[11] H. Bai, H. He, Z. Peng, T. Dai, and S. H. G. Chan. Countr: An end-to-end transformer approach for crowd counting and density estimation. *In European Conference on Computer Vision*, pages 207–222, Cham Switzerland, 2022.

[12] W. Feng, K. Wang, and S. Zhou. An efficient neural network for pig counting and localization by density map estimation. 11: 81079–81091, *IEEE Access*, 2023.

[13] P. Thammasorn, S. Boonchu, and A. Kawewong. Real-time method for counting unseen stacked objects in mobile. *In 2013 IEEE International Conference on Image Processing*, pages 4103–4107, Melbourne, VIC, Australia, 2013.

[14] M. F. Ghazali, L. K. Wong, and J. See. Automatic detection and counting of circular and rectangular steel bars. *In 9th International Conference on Robotic, Vision, Signal Processing and Power Applications: Empowering Research and Innovation*, pages 199–207, Singapore, 2017.

[15] C. Liu, L. Zhu, and X. Zhang. Bundled round bars counting based on iteratively trained SVM. *In Intelligent Computing Theories and Application: 15th International Conference*, pages 3–6, Nanchang, China, 2019.

[16] S. Wang and J. Han, Automated detection of exterior cladding material in urban area from street view images using deep learning, *Journal of Building Engineering*, 96: 110466, 2024

[17] Z. Fan, J. Lu, B. Qiu, T. Jiang, K. An, A. N. Josephraj, and C. Wei. Automated steel bar counting and center localization with convolutional neural networks, *arXiv preprint*, 1906.00891, 2019.

[18] R.-L. Hong, S.-W. Chung, Y. Beak, PetroPipe, (n.d.).https://doi.org/10.34740/KAGGLE/DSV/10939352, Accessed: 03/10/2025.

[19] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 77:157–173, 2008.

[20] S. Wang, Evaluation of impact of image augmentation techniques on two tasks: Window detection and window states detection, *Results in Engineering*, 24: 103571, 2024.