Automatic Person Removal Pipeline

E6691.2023Spring.Advanced_deep_learning.report

Xin Bu
GSAS, Columbia University
Manhattan, USA
xb2165@columbia.edu

Yufan Luo
SEAS, Columbia University
Manhattan, USA
yl5086@columbia.edu

Yuliang Jiang
SEAS, Columbia University
Manhattan, USA
yj2732@columbia.edu

Abstract*—The process of removing special areas from an image and repairing the missing part is known as image inpainting. In this project, the proposed inpainting pipeline can detect persons from the image, and remove them with inpainting. The pipeline is constructed upon state-of-the-art detection, segmentation, and generative models. Later, we will provide the conclusion and future scope of our work. *Keywords—person removal; image inpainting; YOLO; SAM (Segment-Anything Model); image captioning; Stable Diffusion.

I. Introduction

Currently, person removal has been supported by a range of image editing software [1]. However, most of them require masking done by hand, and a lot of them are paid services. At the same time, advanced AI tools in detection and generation have made it possible for us to build such a pipeline ourselves.

In this project, instead of training a model from the ground up, we abide to the power of pre-trained models and bridge the gap between advanced detection models, segmentation models, and generative models to propose an automatic pipeline for person removal. The major work is to select which model fits the job best, and how to connect them efficiently.

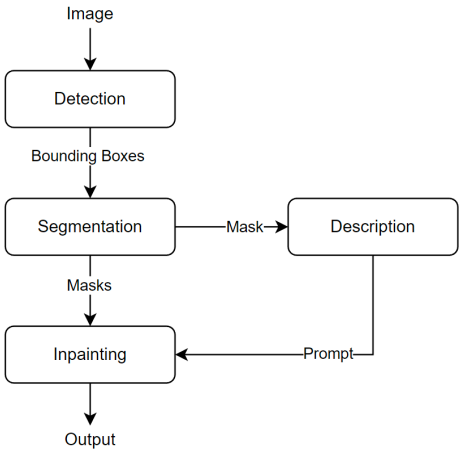

Fig. 1. Out pipeline detects persons with pre-trained detection models

Link to project Github:

https://github.com/Xin-20/e6691-2023spring-project-person_removal

II. Previous Works

Previous work on this theme and area can be categorized into several key approaches: patch-based methods, deep learning-based methods, and hybrid methods.

Patch-based methods, such as the one proposed by Criminisi et al. [2], focus on filling the missing regions in images by searching and copying similar patches from the known parts of the image. Although successful in many cases, patch-based methods often struggle with complex scenes and large objects like persons. Barnes et al. [8] proposed an algorithm called PatchMatch, which significantly improved the efficiency of finding the best matching patches. However, the inherent limitations of patch-based methods remained.

Deep learning-based methods have shown great promise in addressing the challenges faced by patch-based methods. Pathak et al. [9] proposed a method called Context Encoders that used deep convolutional neural networks to fill in missing regions in images. They employed an adversarial loss during training, which further improved inpainting quality. Yu et al. [3] extended this work by introducing a contextual attention mechanism that allowed the model to borrow information from remote regions when reconstructing missing areas. This method demonstrated significant improvements in inpainting quality for complex scenes.

Hybrid methods combine the strengths of both patch-based and deep learning-based methods. For example, the EdgeConnect framework by Nazeri et al. [4] incorporated edge information into the inpainting process, improving the structural coherence of the inpainted regions. This approach combined the use of deep learning with the strengths of patch-based methods to effectively tackle image inpainting tasks.

In terms of person removal, several works have focused on combining object detection and segmentation techniques with image inpainting methods. Liu et al. [10] proposed a method called Detect-Inpaint-Remove (DIR) that integrated object detection, semantic segmentation, and image inpainting to automatically remove unwanted objects from images. Their method used the Mask R-CNN model [11] for object detection and segmentation and employed a deep learning-based inpainting model for object removal.

In summary, the previous work in person removal and image inpainting has evolved from patch-based methods to deep learning-based methods and, more recently, to hybrid methods that leverage the strengths of both approaches. The integration of object detection and segmentation techniques has further enhanced the effectiveness of these methods in removing specific objects, such as persons, from images.

III. METHODS

Traditional image inpainting methods often ignore the presence of people or result in unnatural restoration when dealing with images containing humans. Therefore, our goal is to solve this problem and develop a method that can remove humans from images and restore them naturally. To develop an image inpainting method based on image descriptions and object detection, which can remove humans from images and restore them naturally. We use images from the coco dataset and PennFudanPed for our research.

Two different object detection algorithms, namely, Yolo8, and DETR, are used to detect the positions of humans and generate bounding boxes for SAM to convert them into fine masks. Using gpt2 to generate image descriptions as prompts as input into a stable diffusion model for image inpainting, thus achieving our research goal. Our hypothesis is that using image descriptions and object detection to perform image inpainting can better preserve the authenticity of the original image, remove humans from the image, and inpainting it naturally. We conducted experiments comparing the results of using quick masks directly transformed from bounding boxes of two different object detection algorithms, Yolo8, and DETR, and also compare the result from the fine mask gained by SAM by utilizing the bounding boxes generated by Yolo8 and DETR, to verify our hypothesis and determine the best object detection algorithm. In this part, we will introduce the method we use in this project.

A. YOLO

In this paper, we utilize YOLO v8, an advanced version of the YOLO (You Only Look Once) object detection system, for the person detection task. YOLO, originally introduced by Redmon et al. [5], is a real-time object detection system that has undergone multiple iterations and improvements. YOLO is known for its ability to achieve a balance between high detection accuracy and fast processing speeds, making it suitable for various applications, including real-time image and video analysis.

Our proposed method involves using YOLO v8 to generate bounding boxes around detected persons in the input image. These bounding boxes are then converted to white masks and overlaid on the original image. To make it better, we input bounding boxes to SAM and generate the fine mask, and then converted them to white masks and overlaid them on the original image. The purpose of this step is to emphasize the areas containing persons, allowing the subsequent image captioning model to focus on these regions.

Once the white-masked image is obtained, we employ a GPT-2-based model to generate captions for the modified image. GPT-2, developed by OpenAI, is an autoregressive transformer model known for its strong language modeling and text generation capabilities. By leveraging GPT-2, we aim to generate accurate and contextually relevant captions for the image with masked persons.

After generating the captions, we append the word "background" to the generated captions. This additional step helps to further emphasize the context in which the persons appear, effectively improving the overall performance of the inpainting process. Stable diffusion is used for inpainting with either input of the mask converted by the bounding boxes or input of the find mask generated by SAM with the bounding boxes. By combining YOLO v8 for person detection, white masking for emphasis, GPT-2 for caption generation, SAM for fine mask generation, and the added "background" keyword, our method offers a robust and efficient solution for person removal in images.

B. Segment Anything Model (SAM)

The Segment Anything Model (SAM) is designed to be versatile and efficient in generating object masks for various segmentation tasks. In our project, we provide a more detailed comparison of bounding boxes of DETR and YOLO v8 as SAM prompt input to better understand their performance in person removal and image inpainting applications.

SAM uses a lightweight mask decoder that can be exported to ONNX format, allowing it to run in any environment that supports ONNX runtime, including in-browser applications. The ONNX export feature demonstrates the flexibility of SAM and its potential for integration into various platforms and applications.

The underlying architecture of SAM relies on the Vision Transformer (ViT) backbone, available in different sizes: ViT-H, ViT-L, and ViT-B. These variations allow for trade-offs between computational complexity and model performance, catering to different hardware constraints and application requirements.

In our comparison, we examine the quality of the fine masks generated by SAM and quick masks directly transformed from YOLO v8 or DTER, focusing on the accuracy and precision of the person removal process. This involves assessing the models' ability to handle complex scenes with multiple persons, varying lighting conditions, and occlusions. Furthermore, we evaluate the models' robustness in generating masks for persons with different poses, clothing, and appearances.

In terms of efficiency, we compare the processing time taken by SAM and YOLO v8 to generate masks for the same images. This comparison provides insights into the practical implications of using each model in real-time or resource-constrained scenarios.

Another aspect of our comparison is the ease of integration of SAM and YOLO v8 or DETR into the overall inpainting pipeline. We assess the compatibility of the models with other components of the pipeline, such as segmentation and generative models, and their adaptability to different inpainting scenarios.

By providing a detailed comparison of masks of SAM generated by DETR and YOLO v8 in the context of person

removal and image inpainting, we aim to offer valuable insights for researchers and practitioners working in this field. The results obtained from this comparison can guide the selection of appropriate models for specific applications and inform the development of more effective and efficient person removal and inpainting techniques in the future.

C. Stable Diffusion (SD)

Stable Diffusion (SD) is an advanced image inpainting method that leverages diffusion-based techniques for filling in missing or corrupted areas of an image. This approach has gained popularity due to its ability to generate high-quality inpainted images while maintaining the original image's structure, texture, and color information.

The core idea behind SD is to iteratively diffuse pixel values from the known regions of the image into the missing regions. This diffusion process is guided by a set of partial differential equations (PDEs) that describe the flow of pixel values over time. By solving these PDEs, the algorithm is able to propagate the image information into the unknown regions in a controlled and stable manner.

One key advantage of SD over traditional inpainting methods is its ability to preserve important image features, such as edges and textures while filling in the missing regions. This is achieved by incorporating anisotropic diffusion, which allows the algorithm to adapt the diffusion process based on the local image features. As a result, the inpainted regions blend seamlessly with the surrounding areas, producing visually plausible and coherent results.

In our project, we utilize the `StableDiffusionInpaintingPipeline` class provided in the `inpainting.py` script to incorporate SD into our person removal and inpainting pipeline. This class is designed to accept the masks generated by SAM, allowing for efficient integration with the SAM-based segmentation process.

By combining the high-quality masks generated by SAM with the advanced inpainting capabilities of SD, we aim to create a powerful and effective solution for person removal and image inpainting tasks. This integrated approach has the potential to significantly improve the overall performance and robustness of the inpainting process, making it suitable for a wide range of applications in computer vision and image processing.

D. DEtection TRansformer(DETR)

In our project, we have decided to switch from using YOLO to DETR (DEtection TRansformer) for object detection tasks, as it has demonstrated better performance, particularly when detecting overlapping individuals in images. YOLO struggles with detecting people who are close to each other or partially occluded, whereas DETR can effectively handle such situations.

DETR is a novel end-to-end object detection framework that combines the powerful representation learning capabilities of Transformers with a set-based global loss. The main advantage of DETR over traditional object detection models like YOLO is its ability to reason about the relationships between objects in an image. This is achieved through the use of self-attention mechanisms in the Transformer architecture, which allows the model to efficiently capture long-range dependencies and contextual information.

One significant innovation in DETR is its formulation of the object detection problem as a set prediction task, which eliminates the need for anchor boxes or non-maximum suppression (NMS) commonly used in other detection models. Instead, DETR predicts a fixed number of object bounding boxes along with their class probabilities, and then optimizes the model using a set-based global loss. This approach encourages the model to learn a more robust and accurate representation of the objects in the scene.

By incorporating DETR into our person removal and inpainting pipeline, we can take advantage of its superior performance in detecting overlapping or partially occluded people. This improvement in detection accuracy will directly contribute to the overall quality of the inpainting process, as better object masks will be generated for input to the SAM and SD methods.

The integration of DETR into our pipeline showcases the potential of combining state-of-the-art object detection techniques with advanced inpainting methods to create a robust and effective solution for person removal and image inpainting tasks across various applications in computer vision and image processing.

E. Dataset Description

COCO dataset and PennFudanPed images are used for our validation.

The COCO dataset is a widely used dataset for object detection, segmentation, and image captioning tasks, comprising over 330,000 images and over 2.5 million annotations across 80 categories. Each annotation in this dataset has a category ID and a bounding box or segmentation mask. In our project, we only load human object images and annotations from the COCO dataset, using the COCO Python API to parse the annotations and selecting only the images and annotations for the category "person". In this dataset, each annotation is a segmentation mask for human objects. By using this dataset, it can be used to see how our pipeline performs and make comparisons between models.

The PennFudanPed dataset is a pedestrian detection dataset that consists of 170 images with 345 instances of pedestrians. Each image is of size 720x960 and contains one or several pedestrians. The dataset also includes binary masks for each pedestrian instance, where each mask is labeled with a unique color. The goal of the dataset is to detect all pedestrian instances in each image and predict their corresponding bounding boxes. Here we also use it to see how our pipeline performs and make comparisons between models.

The wild pictures from Columbia University taken by us on campus are used for testing the performance of our final pipeline. The wild pictures mean they are taken in non-specific situations or professional setups.

IV. Experiments

In this section, several techniques of person detection, masking, and prompting are experimentally compared.

A. Experiment with detection techniques

In this experiment, YOLO (YOLOv8n) and DETR are compared as person detection techniques, and DETR shows advantages, especially in crowded scenarios.

As stated in Section III, YOLO is a single-stage detector that makes one prediction per anchor (object center). Thus, when multiple persons are blocked by one another (which is quite common in our dataset), YOLO only detects the closest person and completely omits the blocked persons. This will result in incomplete person masks and can allure SD to generate another person back in place.

Fig. 2. Image with overlapping persons (from PennFudan Dataset).

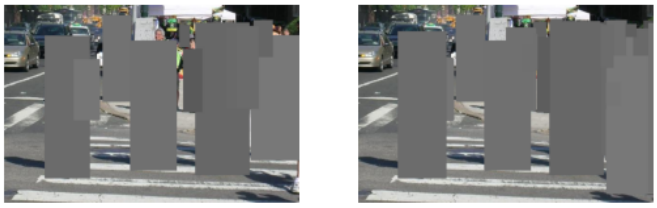

Fig. 3. YOLOv8n (left) omits some persons that can be detected by DETR (right).

Fig. 4. Inpainted image with YOLO (left) and DETR (right).

As shown in Fig. 3, YOLO struggles with severely overlapped persons, especially in the upper right zone. This may not be a serious problem for detection-only tasks, but it is for inpainting. In Fig. 4, we can see SD generates persons back in place due to the remaining pixels of persons.

Additionally, the average time consumption for DETR and YOLO detection over the PennFudan Dataset is shown in Table. I. Although the time difference is prominent, it is still very beneficial to choose DETR over YOLO to detect persons. In practice, some other objects that are closely related to persons, such as handbags, are also on the list for detection, so as to completely remove all pixels that are related to persons.

TABLE I.

Detection Model	YOLOv8n	DETR
Average Time Consumption (ms)	104.03	202.51

B. Experiment with masking techniques

In this experiment, box masks (directly from bounding boxes) and fine masks (created by SAM based on bounding boxes) are compared.

In most scenes, SAM masks block fewer areas, providing more details and thus reducing the chance of generating abrupt objects. In Fig. 7, SD generates some irrelevant objects within the box masks, while it is less likely for SAM masks.

Fig. 5. Uncrowded scene with few persons (from PennFudan Dataset).

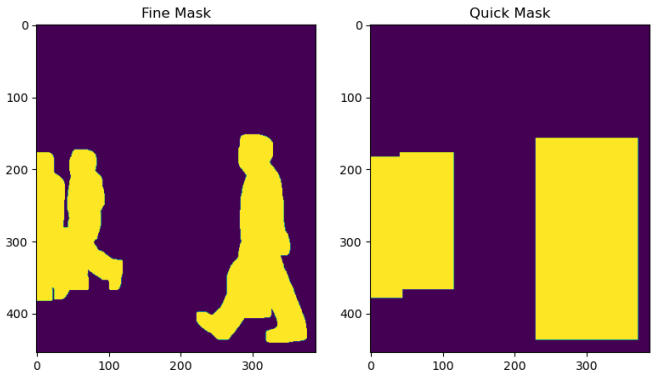

Fig. 6. SAM masks (left) and box masks (right).

Fig. 7. Inpainted image with SAM masks (left) and box masks (right).

However, in the most crowded scenes, SAM masks may create incomplete person masks, giving box masks a better performance. In Fig. 8, we can see shoes appear between the legs of the man in black on the right side. These shoes are left out by SAM mask, leading to undesired results.

Fig. 8. Crowded scene with a lot of persons (from PennFudan Dataset).

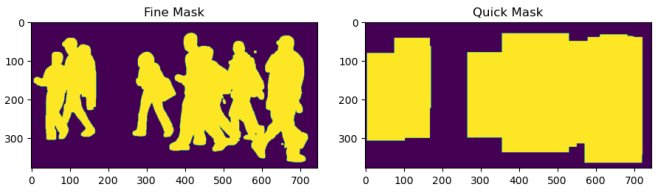

Fig. 9. SAM masks (left) and box masks (right).

Fig. 10. Inpainted images with SAM masks (left) and box masks (right)

Lastly, in terms of time consumption, box mask is the clear winner as shown in Table. II. Given the advantages of neither box masks nor SAM masks are dominating, we keep both masks available and let users decide which mask to use.

TABLE II.

Mask Type	box masks	SAM masks
Average Time Consumption (ms)	44.36	1000.71

C. *Experiment with prompts*

In this experiment, fixed prompts "background", and generative prompts are compared, but their advantages are not clear.

As stated in Section III, GPT-2 is used to generate a caption of the masked image (where all persons are covered by gray masks). The caption is then passed through NLTK (Natural Language Toolkit) with all stopping words removed to form the prompt for Stable Diffusion. However, due to a lack of ground truth, the effect of different prompts cannot be scored and analyzed in an accurate way, but we do observe both types of prompts works better when the person mask is of higher quality.

In terms of time consumption (here refer to prompt generation + inpainting time), fixed prompt is the clear winner as shown in Table. III. Since the effects of different prompts are not clearly explored, we keep both prompt types available and let users decide which prompt to use.

TABLE III.

Prompt Type	fixed	generated
Average Time Consumption (s)	3.55	14.32

V. Discussions

In this project, we successfully constructed an image impainting pipeline to automatically detect and remove persons by comparing and selecting state-of-the-art pretrained AI models.

On one hand, we compared two mainstream person detection algorithms: YOLOv8n and DETR, and found irreplacable advantages for DETR. Also, we discovered that the key to good person removal performance lies in the elimination of person-related pixels (which is to say: high-quality masks at the pixel level), rather than good prompts.

On the other hand, the major drawback is we failed to examine the quality of our generated images in an analytical way due to a lack of ground truth as well as a scoring system (we only examine the quality by manned evaluating).

VI. Results

In comparison to the Stable Diffusion online demo [12], our implementation detects persons more accurately with DETR, providing a better starting point for SD to do inpainting, and thus can remove persons from the image more efficiently.

Fig. 11. Sample images taken in the wild.

Fig. 12. Inpainted images with Stable Diffusion online demo leave abrupt objects.

Fig. 13. Inpainted images with our pipeline look natural.

VII. Future Works

Although our pipeline achieved some advantages over pure Stable Diffusion, there is still one thing as incomplete person masks, which is shadow. There are several papers that related to shadow removal we can apply in future work. The Method of Chromacity information proposed by Sanin et al.[13] can be used to create the mask of candidates' shadow pixels so we can remove them for better performance. The novel deep learning method proposed by Le and Samaras [14] can model the shadow effects in the image by the named SP-Net and M-Net and generate a highly accurate mask of shadow.

Future work for our stable diffusion inpainting model can also focus on improving its ability by fine-tuning the model on our self-designed training set. While our model currently uses neighboring pixels to inpainting missing information, it has no exact too idea about what to fill in, it may still produce images with unwanted objects due to its lack of specific training on person removal.

To address this issue, we suggest designing a specialized dataset for training the inpainting model. The dataset can include pairs of images, where one image contains the expected content and the other image shows the same content with the object to be removed. This will allow the model to learn specifically how to remove objects and produce images with the desired content.

However, collecting such a dataset can be challenging. One approach could be to take photographs of scenes with people and wait for them to leave before taking a second photograph of the same scene. This can be done by using a fixed camera position and timing the photos accordingly. While this method may not be completely accurate since the scene also changes over time, it can provide a starting point for the dataset.

For more professional and controlled images, a better approach would be to use simulation techniques such as ray tracing and 3D modeling proposed by Kolker et al. [15] to create a dataset with images of people in various positions and scenarios. This would require a more complex and time-consuming process, but it could provide a more accurate and diverse dataset for training the model.

With a specialized dataset for object removal, our inpainting model can move beyond surrogate tasks, and achieve end-to-end object removal. We believe this is an important direction for future work, as it can expand the model's capabilities and potential applications.

VIII. Conclusion

We used PennFudanPed and COCO datasets as our validation sets to compare and select models and used our own photographs taken at Columbia University as our testing set. Our results showed that DETR accurately detected people in the images, and our pipeline effectively removed the people from the images and repaired them in a natural way.

Our study find out that by using Yolo/DETR to generate bounding boxes for people, DETR performed better due to its ability to detect multiple overlapping people. We then input the generated bounding boxes into SAM to generate person segmentation masks, which we converted to white masks and used as input to gpt2 to generate a background description. We used this description as a prompt in stable diffusion for image inpainting, which resulted in successful and natural-looking image repairs. We also experimented with using YOLOv8n/DETR-generated bounding boxes directly without using SAM for person segmentation, converting the person bounding boxes to white masks and inputting them to gpt2 to generate a background description, and using this description as a prompt in stable diffusion for image inpainting. However, we found that this approach resulted in a rectangular blur region that was not as effective as using SAM-generated masks as input for stable diffusion.

Acknowledgments

We would like to express our deepest appreciation to Professor Mehmet Turkcan for his dedicated teaching in the course *Advanced Deep Learning*, and StabilityAI for their thorough research in the stable diffusion architecture which motivated us to build upon their work.

References

[1] N. Shan, D. S. Tan, M. S. Denekew, Y. -Y. Chen, W. -H. Cheng and K. -L. Hua, "Photobomb Defusal Expert: Automatically Remove Distracting People From Photos," in IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 4, no. 5, pp. 717-727, Oct. 2020, doi: 10.1109/TETCI.2018.2865215.

[2] Criminisi, A., Pérez, P., & Toyama, K. (2004). Region filling and object removal by exemplar-based image inpainting. IEEE Transactions on Image Processing, 13(9), 1200-1212.

[3] Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S. (2018). Generative Image Inpainting with Contextual Attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5505-5514).

[4] Nazeri, K., Ng, E., & Ebrahimi, M. (2019). EdgeConnect: Structure Guided Image Inpainting using Edge Prediction. In Proceedings of the IEEE International Conference on Image Processing (pp. 2221-2225)

[5] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 779-788).

[6] GitHubRepository: Segment-AnythingModel(SAM) https://github.com/segment-anything/SAM

[7] GitHubRepository: DeepFill https://github.com/JiahuiYu/generative_inpainting

[8] Barnes, C., Shechtman, E., Finkelstein, A., & Goldman, D. B. (2009). PatchMatch: A Randomized Correspondence Algorithm for Structural Image Editing. ACM Transactions on Graphics (TOG), 28(3), 1-11.

[9] Pathak, D., Krahenb uhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context Encoders: Feature Learning by Inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2536-2544).

[10] Liu, Z., Li, X., Luo, P., Loy, C. C., & Tang, X. (2018). Deep Learning Markov Random Field for Semantic Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(8), 1814-1828.

[11] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2961-2969).

[12] Stable Diffusion Multi Inpainting, https://huggingface.co/spaces/multimodalart/stable-diffusion-inpainting

[13] A. Sanin, C. Sanderson and B. C. Lovell, “Improved Shadow Removal for Robust Person Tracking in Surveillance Scenarios,” 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 2010, pp. 141-144, doi: 10.1109/ICPR.2010.43.

[14] Hieu Le, Dimitris Samaras, “Shadow Removal via Shadow Image Decomposition”

[15] Kolker, A., Oshchepkova, S., Pershina, Z., Dimitrov, L., Ivanov, V., Rashid, A., & Bdiwi, M. (2020). The ray tracing based tool for generation artificial images and neural network training. Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. https://doi.org/10.5220/0010168102570264