An Innovative Zero-Shot Inference Approach Based on Deep Learning

Zhuo Lei[*, 1, 2], Wei Li[*, 1, 2], Xiangwei Zhang[1], Qiang Yu[1], Lidan Shou[2], Shengquan Li[1] , Yunqing Mao [1]

[1]College of Computer Science and Technology, Zhejiang University, Hangzhou, 310000, China
[2]City Cloud Technology(China) Co., Ltd., Hangzhou, 310000, China
leizhuo@citycloud.com.cn

Abstract. We present a novel zero-shot inference framework for urban management applications, particularly in retail environments. The deep learning model fuses multi-scale CNN-based object detection with self-attention mechanisms to enhance the identification of unauthorized activities and complex categorization tasks in fixed-point surveillance scenarios. Innovative components include lightweight channel aggregation modules that reduce high-dimensional representations and intermediate interactions are captured through multi-stage gate aggregation. Spatial aggregation extracts context-aware multi-level features, addressing limitations of traditional DNN. Attention down-sampling is integrated to address computational challenges when applying Transformers on high-resolution imagery. Explicit labels are trained on raw text-image pairs using contrastive learning. This enables the model to learn from natural language supervision and perform zero-shot recognition across unseen categories. We obtain the state-of-the-art performance both in public dataset and our own urban management dataset.

Keywords: Zero-shot Learning, Object Detection, Transformer

1 Introduction

In the past, urban management heavily relied on time-consuming and inefficient manual procedures. However, recent technological advancements, particularly in the realm of computer vision, have now empowered a more effective response to these challenges. Leveraging real-time monitoring, and analytical capabilities enables swift detection of infractions, comprehensive environmental surveillance, as well as strategic allocation of resources, thereby significantly enhancing the precision and efficiency of urban administration. In the current era, big data's inherent diversity has given rise to

[*] The first author and the second contribute equally.

the prominence of text and image data as an abundant resource.In modern urban management systems, there is a regulatory environment with zero or few samples to learn, and urban management departments heavily rely on manual supervision that is complex and inefficient.

We propose a deep learning-based zero-shot inference approach designed to enhance intelligent detection of store management objectives through the use of fixed and mobile cameras, thereby streamlining urban management operations. Specifically, our aim is to apply advanced deep learning techniques in retail environments for the identification of goods. While a multitude of solutions (e.g., [1, 2]) have been introduced to expedite attention mechanisms, their practical application can encounter limitations. Since the groundbreaking success of Transformer [3] in NLP tasks as evidenced by [4, 5], ViT [6] has been introduced and demonstrated promising outcomes on the ImageNet benchmark [7]. However, compared to ConvNets, pure ViTs are generally more over-parameterized and heavily reliant on large-scale pre-training datasets [8, 9, 2]. To tackle this issue, one line of research has focused on proposing lightweight ViT architectures [10, 11, 12], often incorporating efficient attention mechanisms [13]. Concurrently, there has been an active exploration into hybrid backbones that fuse self-attention with convolution [14, 15, 16, 17] to imbue ViTs with region-specific priors typical of ConvNets. Through the integration of inductive biases [18, 19, 20], or supplementary knowledge [21], ViT and its derivatives have managed to achieve competitive performance levels akin to ConvNets. Improving the efficiency of Transformers without compromising effectiveness remains a significant research topic, with several design and usage challenges yet to be fully addressed. Our key technical innovations consist of:

1) Employing a lightweight channel aggregation module that re-weights high-dimensional hidden spaces and diminishes projection channels.

2) Integrating multi-stage gate aggregation modules to capture comprehensive intermediate interactions.

3) Down-sampling to a uniform spatial resolution prior to interpolating the attention output back to its original size before feeding it into subsequent layers.

4) Incorporating an attention down-sampling module, to form a hybrid local-global structure to bolster performance.

We obtain the state-of-the-art performance both in public dataset and our own urban management dataset.

2 The Proposed Framework

We propose an innovative algorithm that combines a CNN based target detection framework with a self-attention mechanism within a zero-shot inference architecture. Multi-scale design allows the model to capture object feature at different resolutions, thereby better adapting to changes in different size and illegal activity detection. The self-attention mechanism dynamically weights different regions, enabling the model to automatically focus on the most critical parts. The combination of the two not only enhances the model's ability to identify subtle and diverse violations in fixed point monitoring, but also promotes zero-shot learning.This system is primarily tailored for fixed-point surveillance to identify unauthorized out-of-store business activities and perform intricate text and image categorization.

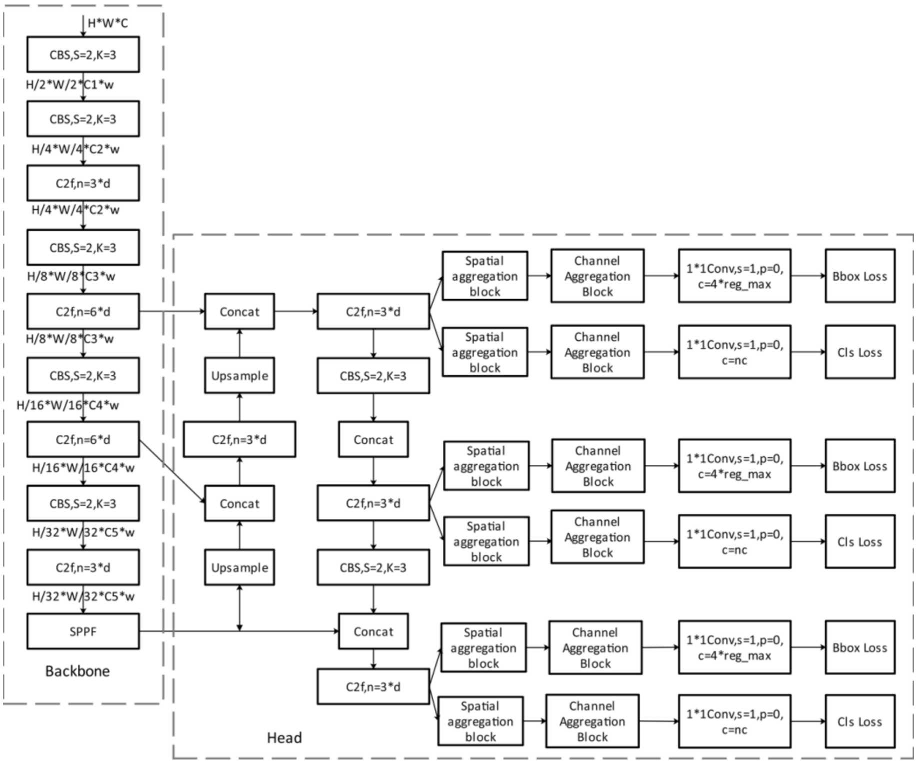

Fig. 1. The multi-scale CNN architecture. Within the spatial aggregation module, a multi-order gating aggregation sub-module effectively captures multi-level interactions within the context, extracting multi-level features that exhibit both static and adaptive regional sensitivity.

2.1 Network Design

Multi-scale convolutional neural network. As depicted in Figure 1, the framework is divided into the backbone network and the head network. The backbone network undergoes five down-sampling stages where the input label image is resized to a dimension of $H \times W$, featuring a cross-stage partial network design. To further enhance its lightweight nature, we incorporate the $C2f$ and SPPF. The head network adopts the Feature Pyramid Path Enhancement Network, which functions as a decoupled header component. It utilizes a loss function that combines BCE Loss for classification purposes, while using the VFL and the CIOU Loss for regression. Our framework employs a Task Alignment Allocator matching method based on an Anchor-Free approach. The parameter w governs the width of the network, while d controls the depth of the network structure.

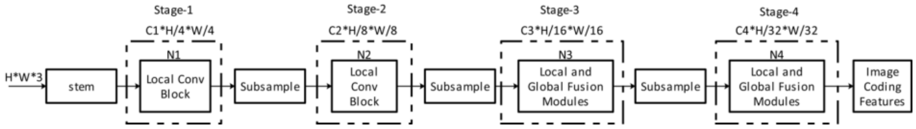

Fig. 2. Convolutional attention mechanism can be added to enhance the performance and performance of the model.

In Figure 2, the $C2f$ module consists of a CBS convolution module, a channel splitting operation, n residual bottleneck modules, branch stacking and a CBS feature convolution and channel reduction. The residual bottleneck module is depicted in detail, where the upper branch consists of a sequential arrangement of two CBS modules. Both these CBS modules employ a kernel size of 3×3, a stride of 1, and a padding of 1 for their convolutional layers. The first CBS within this sequence serves to reduce the feature dimensions, while the second subsequently increases the dimensions back. Regarding the bypass link, the presence or absence of the component is controlled by the $Shortcut$parameter. If $Shortcut = True$, it signifies that a bypass link operation is included, which allows the original input features to be added directly to the output of the two CBS modules. Conversely, when $Shortcut = False$, no such bypass connection is employed, meaning that only the processed features from the CBS modules are passed on to subsequent layers.

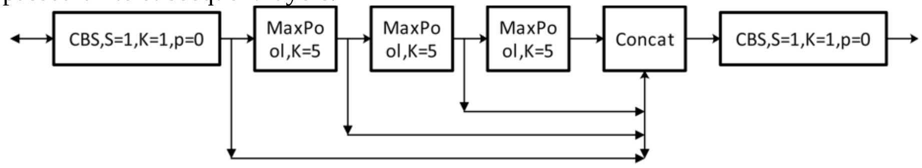

Fig. 3. The SPPF module. It divides receptive fields of the same size into different levels and performs pooling operations on the features within each level, thus enabling fixed length feature representation of input images of any size.

The SPPF is illustrated in Figure 3, serving the purpose of achieving adaptive output sizes. In SPPF, as depicted, it commences with a CBS module where C stands for convolutional operation, B represents batch normalization, and S denotes the $SiLU$ activation function. Here, $K = 1$ indicates a kernel size of 1×1, $S = 1$implies a stride of 1, and $p = 0$means there is no padding around the feature map. Subsequently, three consecutive max pooling layers are connected, each with a pooling kernel size specified by $K = 5$. Thirdly, there's a feature stacking phase where the output results of these three max pooling layers are concatenated or stacked together with the output from the initial CBS module. This fusion of different scale features enables the model to capture context information at multiple levels. Finally, the stacked features undergo another CBS feature extraction operation.

Figure 4 illustrates the Spatial Aggregation Module, which is designed to address the limitation of traditional deep neural networks. This module efficiently captures multi-order contextual interactions by extracting features with both static and adaptive region awareness. In this process, the 0-order interaction inherent in each patch and the 1-order interaction encompassing all patches are modeled using 1×1 convolutions and global average pooling. To enforce the network's focus on multi-order interactions, less significant interaction components are dynamically down-weighted. Firstly, a 1×1 convolution operation is applied to obtain feature map $X1$. Global average pooling is then conducted on $X1$ to generate $X2$. By subtracting $X2$ from $X1$, we derive feature Y. A dot product between Y and its scaled version Ys (where Ysrepresents a scaling factor) produces feature U. Features $X1$and $U$are combined through element-wise addition to yield$Z1$, followed by a non-linear activation function $SiLU$ to produce the final feature Z. The subtraction step $X1 - X2$ serves to re-weight unimportant interaction components, thereby enhancing feature diversity. Subsequently, DWConv is integrated into the context branch of the multi-order aggregation module to encode multi-order features. This module utilizes n depth convolution layers in a progressive manner to capture low-, middle- and high-order interactions, ensuring a comprehensive understanding of the multi-scale context.

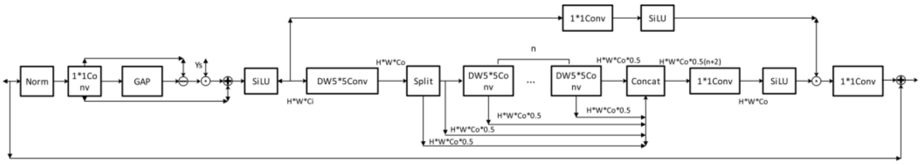

Fig. 4. The spatial aggregation module effectively captures multi-level interactions in the context, extracting multi-level features with static and adaptive region perception.

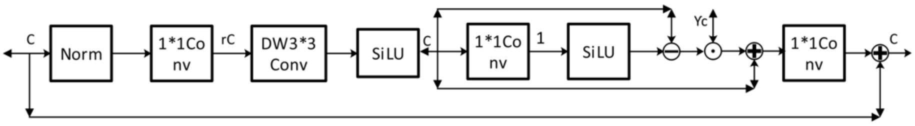

Fig. 5. A lightweight channel aggregation module is used to re-weight high-dimensional hidden spaces, to collect and reallocate channel level information by reducing projection channel features and activation functions.

In Figure 4, the $SiLU$ activation function is employed in the gating branch. This choice leverages the gating properties of $Sigmoid$, which can selectively pass through or suppress information, while also maintaining the stable training characteristics that help prevent issues. The multi-order gating aggregation module specifically targets capturing a richer set of middle-order interactions, which are often overlooked in conventional approaches. However, achieving the desired performance level may necessitate a larger multi-layer perceptron ratio, which could introduce additional parameters and computational overhead, potentially due to redundancy across channels. To address this issue, as illustrated in Figure 5, a Lightweight Channel Aggregation Module is utilized. It serves to re-weight the high-dimensional hidden space by reducing the projection channel and applying activation functions.

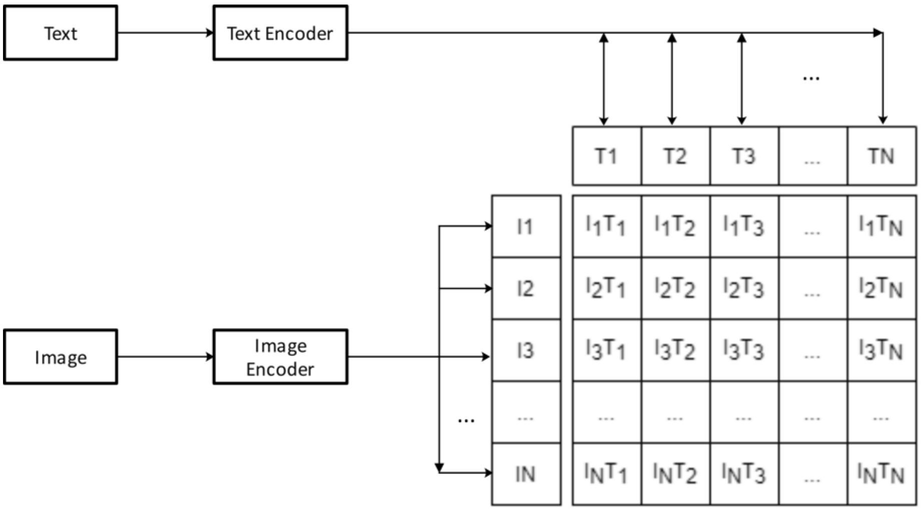
Text
Text Encoder
...
T1
T2
T3
...
TN
Image
Image Encoder
I1
I2
I3
...
IN
I_1T_1 I_1T_2 I_1T_3 ... I_1T_N
I_2T_1 I_2T_2 I_2T_3 ... I_2T_N
I_3T_1 I_3T_2 I_3T_3 ... I_3T_N
... ...
I_NT_1 I_NT_2 I_NT_3 ... I_NT_N

Fig. 6. The multi-modal network architecture includes image encoding and text encoding.

Multi-modal Model Construction. Current computer vision models are conventionally trained to recognize and classify a predetermined, fixed set of classes. For instance, like ImageNet (with 1000 predefined classes)[7] or COCO (with 80 classes) [22], this form of supervision is restrictive because it requires substantial amounts of labeled data and limits the model's ability to generalize to unseen categories. An alternative and increasingly promising approach is to train models directly from raw text associated with images, bypassing the need for explicit class labels. In such a scenario, a model can learn by predicting which captions correspond to which images, serving as an effective and scalable pre-training strategy. First, it does not necessitate manually labeling large volumes of data. The inherent structure and semantics within the text provide a rich source of supervision that can grow organically alongside the available textual information about images on the web or other sources. Second, by connecting visual representations with linguistic descriptions, the model gains the capability to perform zero-shot transfer-meaning it can potentially identify objects or scenes from novel classes. The connection between the visual and textual modalities allows the model to understand and reason about new concepts based on its existing knowledge. Finally, training image and text embedding jointly results in a learned multi-modal representation that can more readily adapt to diverse tasks and scenarios.

The multi-modal training architecture described here is designed to process and fuse information from both visual and textual inputs for a joint understanding of the data. The architecture accepts two types of inputs. Image input has dimensions $[n, h, w, c]$, where n refers to the batch size, h and w denote the height and width of each image, and c represents the number of color channels. Textual descriptions come in dimensions $[n,1]$, with n again being the batch size and 1 representing the variable sequence length of words or tokens. Each input type is passed through its own encoding module. For images, this could involve CNN to extract high-level visual features. For text, it would typically involve a transformer-based model or RNN to generate contextualized word embedding. After extracting uni-modal features, there are projection layers denoted by Wi for image features and Wt for text features. These projections adaptively transform each modality into a shared multi-modal space where they can be compared effectively. The projected features (Ie and Te for images and text, respectively) undergo $L2$ normalization before being compared to measure their similarity or correlation. The depth N, width C, and feed-forward network expansion rate of the overall architecture are determined using NAS techniques. This approach automates the design process to find an optimal configuration that balances efficiency and performance. During the NAS process, the number of $MHAS$ blocks is dynamically adjusted. In the last stages of the network, every block consists of a $MHAS$ layer followed by a FFN. The search decides how many global $MHAS$ modules should be retained, ensuring the best trade-off between computational complexity and feature extraction capability. $SiLU$ is used instead of $GELU$ due to its comparable performance but potentially better storage efficiency. The initial part of the network consists of two 3×3convolution operations with a stride of 2and padding of 1 to down-sample the input images while preserving spatial information.

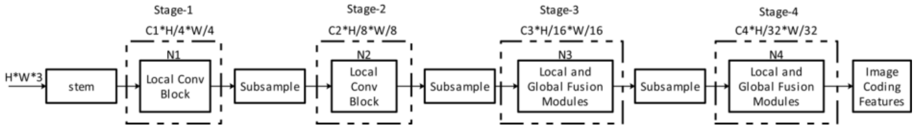

Fig. 7. The image encoding module adopts both convolutional local module and local and global context combination module.

In Figure 8, the Convolutional Local Module is presented in a residual network structure. This architecture allows for efficient and effective learning by preserving the identity of input features while refining them through convolutional operations. The upper branch of this module consists of a series of convolutional layers, normalization steps, and nonlinear activation functions. These components work together to extract and refine local features from the input data. On the other hand, the lower branch serves as a bypass connection that directly forwards the input features to the output without any transformation. This shortcut path maintains the original information and helps prevent the vanishing gradient problem during training. The outputs of both branches are then combined through an element-wise addition operation. By doing so, the module leverages the benefits of both the refined local features from the convolutional pathway and the preserved raw features from the shortcut connection. This design effectively harnesses the inductive bias inherent in convolutional operations while allowing the model to learn residuals between the input and its more complex representation.

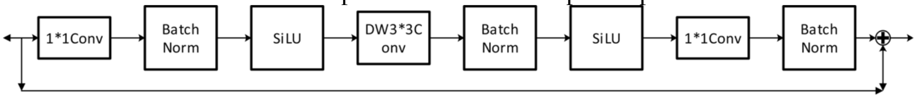

Fig. 8. The inductive bias in the convolutional local module is beneficial for the model to quickly converge from a large number of latent function spaces to subsets.

The attention mechanism has also proven beneficial for computer vision tasks where it can enhance feature representation by selectively focusing on salient parts of images. However, a challenge arises when applying attention to high-resolution feature maps because its computational complexity grows quadratically with respect to the spatial dimensions. This is due to the dot-product attention operation that compares every element across the entire space. In Figure 9, $Query$, Key, and $Value$ matrices are down-sampled or sub-sampled to a lower fixed spatial resolution before performing the attention calculation. This reduction in size significantly reduces the computational load while still capturing essential information from the original high-resolution features. After the attention mechanism processes these lower-resolution representations, the output is up-sampled or interpolated back to match the original input's resolution. The purpose is to ensure that the refined attention-based features can be seamlessly integrated into the subsequent layers of the network, maintaining the consistency and integrity of the feature hierarchy.

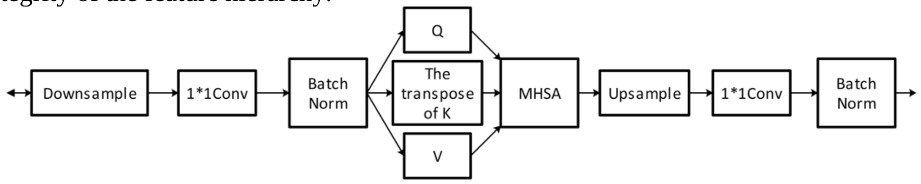

Fig. 9. The combination of local and global context modules effectively reduces the quadratic complexity of the transformer.

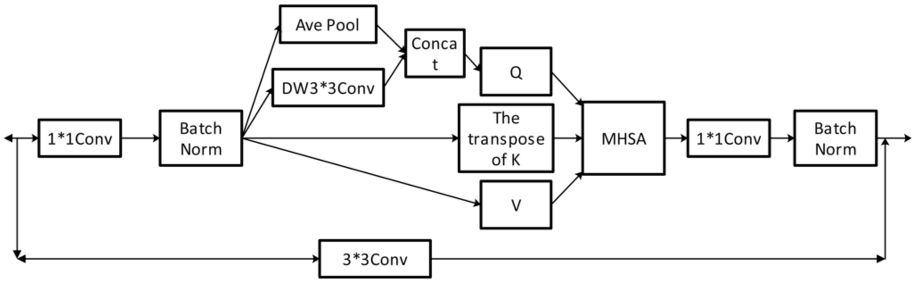

Fig. 10. We obtain the $Query$of down-sampling, using pooling layers as static local down-sampling, using 3×3 depth-wise separable convolutions as learnable local down-sampling, and concatenating the results and projecting them into the $Query$ matrix.

In Figure 10, the process of obtaining a sub-sampled $Query$ representation is depicted to improve computational efficiency while retaining key information in the attention mechanism.To achieve this, the module employs two types of sub-sampling. Static local sub-sampling involves using a pooling layer to down-sample the input features spatially by summarizing local information into a lower resolution. Pooling layers are fixed functions and do not have learnable parameters. Learnable local sub-sampling means a 3×3 DWCONV is used for this purpose. DWCONVs apply a single filter per input channel, allowing for more efficient feature extraction with fewer parameters compared to standard convolutions. The learned filters can adaptively capture important local patterns in the data, thus providing a learn able alternative to the static sub-sampling. The results from these two sub-sampling methods are concatenated together, effectively merging the strengths of both static and learnable down-sampling techniques. Following concatenation, a projection operation is performed to transform this combined information into the desired $Query$ matrix format. Moreover, this Attention Sub-sampling Module is designed as a residual connection to a convolutional layer with a stride. This configuration creates a local-global interaction mode where the sub-sampled features interact with those captured by the strided convolution.

3 EXPERIMENTS

3.1 Datasets

We use images captured by mobile phones, which are divided into distant and close-up target images. Moreover, we also use images from the monitoring videos. From monitoring videos, L distinct segments Vi are identified. This diversity is crucial for ensuring that the model being trained can generalize well across various situations. Each segment Vi likely contains one or more instances of these target objects. Within every video segment Vi, there are Ni total number of frames. Mi images are selected from each segment. By pooling together Mi images from each of the L segments, the overall dataset for training and testing comprises $\sum_{L}^{i=1} Mi$ images. There are 90925 mobile phone images and 77491collected monitoring video frames. Meanwhile, we also use the public COCO dataset to conduct the experiments.

3.2 Experiment Configuration

The training configuration is basically consistent from the baseline model to the final model, training for 300 epochs on our own data and performing 5 epochs of warm-up. The optimizer is SGD. The learning rate is, the initial learning rate is 0.01. Cosine learning mechanism is used. Weight decay is set to 0.0005. Momentum is set to 0.90. Batch depends on the hardware device. Input size transitions uniformly from 448 to $832$with a step size of 32. We randomly initialize the connection weights w and biases b of each layer. Given the learning rate η and the minimum batch $Batch$, we select the activation function $SMU$and the maximum number of iterations. During model training, multiple GPUs are used when the hardware meets the requirements, the deep learning framework used for training is PyTorch.

3.3 RESULTS AND DISCUSSION

Table 1 summarizes the performance of our methods and contrasts to those attained by prior work. It indicate that the proposed method obtains the best performance in the corresponding setting. We achieve the highest mean Average precison scores on both mobile phone image dataset and Monitoring Video dataset. Moreover, we achieve comparable results with other state-of-the-art YOLO methods on the COCO dataset [22]. The results demonstrate that our method effectively captures image semantics in the context of urban management issues. This indicates that the lightweight channel aggregation module efficiently weights the high-dimensional latent space, thereby reducing the need for projection channel features and activation function implementations, thus enabling a more effective gathering and redistribution of channel-wise information. Concurrently, the multi-stage gated aggregation modules capture a greater amount of intermediate-level interactions. On another front, down-sampling to a fixed spatial resolution followed by interpolation of attention outputs back to the original resolution before feeding them into the subsequent layer, also serves to decrease both floating-point operations and parameter counts. Lastly, the attention-based down-sampling module is residually connected to a strided convolution, forming a local-global pattern, which enhances the semantic representation of images in cases pertaining to unauthorized street vending problems.

Table 1. Performance on the three datasets. We compare our approach with the state-of-the-art methods (YOLOs), and we obtain the best results. The best results are highlighted.

	model	YOLOX -S	PPYOLOE -S	YOLOv5 -S(6.1)	YOLOR -CSR	YOLOv7	YOLOv5 -S6(6.1)	YOLOR -P6	Ours
	#Para	9.0M	7.9M	7.2M	52.9M	36.9M	12.6M	37.2M	7.9M
	FLOPS	26.8G	17.4G	16.5G	120.4G	104.7G	67.2G	325.6G	17.9G
	Size	640	640	640	640	640	1280	1280	640
	FPS	102	208	156	106	161	122	76	149
COCO	AP^{test}/ AP^{val}	40.5%/ 40.5%	43.1%/ 42.7%	-/ 37.4%	51.1%/ 50.8%	51.4%/ 51.2%	-/ 44.8%	53.9%/ 53.5%	**54.0%/ 53.6%**
	$AP^{test}_{0.5}$	-	60.5%	-	69.6%	69.7%	-	**71.4%**	71.1%
	$AP^{test}_{0.75}$	-	46.6%	-	55.7%	55.9%	-	58.9%	**59.1%**
	AP^{test}_{S}	-	23.2%	-	31.7%	31.8%	-	**36.1%**	**36.1%**
	AP^{test}_{M}	-	46.4%	-	55.3%	55.5%	-	57.7%	**57.9%**
	AP^{test}_{L}	-	56.9%	-	64.7%	65.0%	-	65.6%	**65.7%**
Mobile Phone	$mAP^{test}_{0.5}$	65.7%	66.4%	64.1%	67.2%	73.0%	72.8%	73.0%	**73.7%**
	$mAP^{test}_{0.5;0.95}$	34.1%	34.6%	33.3%	35.8%	37.9%	39.9%	39.9%	**40.2%**
Monitoring Video	$mAP^{test}_{0.5}$	62.5%	64.9%	62.5%	69.6%	69.7%	71.5%	71.5%	**71.9%**
	$mAP^{test}_{0.5;0.95}$	31.6%	32.7%	31.8%	35.7%	37.4%	40.0%	42.1%	**42.9%**

4 CONCLUSION

A zero-shot inference method based on deep learning is devised to address the challenge of intelligent detection of out-of-store business activities in urban management scenarios using both fixed cameras and mobile phone footage. This innovative approach utilizes images captured by stationary surveillance cameras installed as part of urban infrastructure, feeding them into an algorithm that automatically detects unauthorized commercial activities within the camera's field of view. The methodology provides a convenient, swift, and transparent platform for urban administrators to manage such information. It harnesses the power of deep learning to enable efficient and effective urban intelligence management and operations. Notably, it archives and records detection of out-of-store businesses under monitoring, enabling verification by the management department. Timely notifications are sent to relevant personnel, prompting them to promptly attend the scene for necessary actions. Meanwhile, we advance object detection by combining state-of-the-art CNN and Transformer-like architectures, enabling efficient and effective urban administration via real-time monitoring and strategic resource allocation. We obtain the state-of-the-art performance both in public dataset and our own urban management dataset.

References

1. Huiqi Deng, Qihan Ren, Hao Zhang, and Quanshi Zhang, “Discovering and explaining the representation bottleneck of DNNS,” in ICLR. 2022, OpenReview.net.
2. Siyuan Li, Di Wu, Fang Wu, Zelin Zang, and Stan Z. Li, “Architecture-agnostic masked image modeling - from vit back to CNN,” in ICML, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, Eds. 2023, vol. 202 of Proceedings of Machine Learning Research, pp. 20149– 20167, PMLR.
3. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in NeurIPS, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, Eds., 2017, pp. 5998– 6008.
4. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Jill Burstein, Christy Doran, and Thamar Solorio, Eds. 2019, pp. 4171–4186, Association for Computational Linguistics.
5. Tom B. Brown, Benjamin Mann, Nick Ryder, MelanieSubbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei, “Language models are few-shot learners,” in NeurIPS, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, Eds., 2020.

6. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in ICLR. 2021, OpenReview.net.
7. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in CVPR. 2009, pp. 248–255, IEEE Computer Society.
8. Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei, "Beit: BERT pre-training of image transformers," in ICLR. 2022, OpenReview.net.
9. Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross B. Girshick, "Masked autoencoders are scalable vision learners," in CVPR. 2022, pp. 15979–15988, IEEE.
10. Sachin Mehta and Mohammad Rastegari, "Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer," in ICLR. 2022, OpenReview.net.
11. Yanyu Li, Geng Yuan, Yang Wen, Ju Hu, Georgios Evangelidis, Sergey Tulyakov, Yanzhi Wang, and Jian Ren, "Efficientformer: Vision transformers at mobilenet speed," in NeurIPS, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, Eds., 2022.
12. Mengzhao Chen, Mingbao Lin, Ke Li, Yunhang Shen, Yongjian Wu, Fei Chao, and Rongrong Ji, "Cf-vit: A general coarse-to-fine method for vision transformer," in AAAI, Brian Williams, Yiling Chen, and Jennifer Neville, Eds. 2023, pp. 7042–7052, AAAI Press.
13. Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma, "Linformer: Self-attention with linear complexity," CoRR, vol. abs/2006.04768, 2020.
14. Jianyuan Guo, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang, and Chang Xu, "CMT: convolutional neural networks meet vision transformers," in CVPR. 2022, pp. 12165–12175, IEEE.
15. Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao, "Uniformer: Unifying convolution and self-attention for visual recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 10, pp. 12581–12600, 2023.
16. Zizheng Pan, Bohan Zhuang, Haoyu He, Jing Liu, and Jianfei Cai, "Less is more: Pay less attention in vision transformers," in AAAI. 2022, pp. 2035–2043, AAAI Press.
17. Chenyang Si, Weihao Yu, Pan Zhou, Yichen Zhou, Xinchao Wang, and Shuicheng Yan, "Inception transformer," in NeurIPS, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, Eds., 2022.
18. Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai, "Deformable DETR: deformable transformers for end-to-end object detection," in ICLR. 2021, OpenReview.net.
19. Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao, "Pre-trained image processing transformer," in CVPR. 2021, pp. 12299–12310, Computer Vision Foundation / IEEE.
20. Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid, "Vivit: A video vision transformer," in ICCV. 2021, pp. 6816–6826, IEEE.
21. Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan, "Tinyvit: Fast pretraining distillation for small vision transformers," in ECCV, Shai Avidan, Gabriel J. Brostow, Moustapha Ciss´e, Giovanni Maria Farinella, and Tal Hassner, Eds. 2022, vol. 13681 of Lecture Notes in Computer Science, pp. 68–85, Springer.
22. Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll´ar, and C. Lawrence Zitnick, "Microsoft COCO: common objects in context," in ECCV, David J. Fleet, Tom´as Pajdla, Bernt Schiele, and Tinne Tuytelaars, Eds. 2014, vol. 8693, pp. 740–755, Springer.

Reviewer 1:

1. The abstract section of the manuscript provides a clear overview of the main contributions. However, the advantages of combining multi-scale target detection and self-attention mechanism are not clearly expressed. It is recommended to add more clarification.

The advantages have been added in the beginning of the proposed framework.

2. The introduction of the manuscript mentions multimodal learning, but does not elaborate on its research status. It is recommended to add a description of its research status.

The multimodal learning is not quite belong to the main contribution, so the related content has been removed mostly.

3. The conclusion of the manuscript contains too little description of future research directions. It is recommended to add future prospects for related work research.

Due to the page limit, the text related to future work has been removed.

4. There are several grammatical errors. The author is advised to check the article and correct grammatical errors.

The grammatical errors has been checked and corrected.

Reviewer 2:

1. Why is the multi-scale CNN in Figure 1 divided into two modules with two dotted boxes, and what are the functions of these two modules?

The functions has been explained in the related part.

2. Redraw Figure 2 to make the diagram clearer.

The figure has been redrew.

3. There are many grammatical errors in the whole text. For example, the title of Figure 3 does not know what it is described.

The description of Figure 3 has been modified.

4. Redraw Figure 4 to make the diagram clearer.

The figure has been redrew.

5. The whole paper is chaotic, and there is no formula to precisely define the specific functions of each module.

The contribution and explanation has been modified in the related part.

Reviewer 3:

1. The clarity of the figures is very low and many of the figures cannot be seen clearly, much less the key parts.

All the figures have been redrew.

2. The problem that the paper is trying to solve is not well interpreted.

The problem part has been modified in the introduction part.

3. The method proposed in the article does not have better experimental results to prove its effectiveness.
We have highlighted the best results in the table, and we obtain the state-of-the-art results.

4. Multi-modal Model Construction Fig. 6 does not seem to be different from CLIP.
Figure 6 may have similarities with CLIP in form, but through targeted design and optimization, our model has demonstrated unique value and performance advantages in solving practical problems in urban management systems.

5. The readability of the whole paper is poor, and the Chinese English is serious.
We will keep improving our English.

Reviewer 4
1. Generally speaking, the present paper's discourse is rather succinct, necessitating a more comprehensive exposition through the incorporation of formulas. Furthermore, the paper's experimental section requires enhancement, as the current iterations are inadequate.
Due to the time and page limit, more experiment result cannot be conducted in time. We will improve the experiments in our future work.