Feature Transfer Block for Feature Fusion in Lightweight Object Detectors

Haoyuan Liu
Graduate School of Fundamental Science and Engineering
Waseda University
Tokyo, Japan
liuhaoyuan@akane.waseda.jp

Hiroshi Watanabe
Graduate School of Fundamental Science and Engineering
Waseda University
Tokyo, Japan
hiroshi.watanabe@waseda.jp

***Abstract*—Modern object detectors often use multi-scale detection to improve their performance. To enrich the semantic information from the original feature map generated by backbone, various pyramid-liked architectures such as FPN and PANet are used as the neck of an object detector. However, most researches focus on the macro design of the topology connection of the pyramid structure while putting less effort into the micro design of feature fusion strategy. To address this, we propose a novel fusion strategy that adopts a cross-attention fashion. We named it Feature Transfer Block (FTB) as it explicitly selects semantic information from different scales of feature maps for fusion. This design is cost-efficient and easy to integrate with current designs. Experimental results demonstrate a moderate improvement of YOLOX series on COCO and CrowdHuman when implemented with FTB.**

***Index Terms*—object detection, feature fusion, cross-attention**

I. Introduction

Object detection is a fundamental and crucial task in computer vision, serving as the foundation for image comprehension and processing. It involves identifying objects of interest in input images or videos and providing essential information about their categories and locations.

Modern object detectors commonly employ multi-scale schemes for detection. By using feature maps from various depths of the feature hierarchy, these models can comprehend semantic information at different scales, enabling predictions for objects of different sizes. Several architectures, such as Feature Pyramid Network(FPN) [1], Path Aggregation Network(PANet) [2], and BiFPN [3], serve as necks for object detectors. Among these, the PANet-like structure is frequently utilized in contemporary YOLO-series designs.

Although various pyramid-like networks have been introduced to enhance the multi-scale feature fusion process as mentioned above, adding more connections between levels may increase the computational cost and lead to diminishing returns in terms of overall improvement, resulting in suboptimal fusion outcomes.

The current fusion strategy, which typically employs simple concatenation or element-wise addition, heavily relies on the computational capacity of adjacent ConvBlocks. While these ConvBlocks are concurrently responsible for feature extraction, the fusion process may fail to effectively merge semantic information from different layers under constrained computation resources.

To tackle this issue, our paper propose a novel feature fusion strategy which allows the detection to have better performance with negligible extra computational cost. The main contributions in this paper can be summarized as follows:

1) We propose the Feature Transfer Block (FTB) to replace traditional feature fusion architectures. FTB allows for merging semantic features by explicitly selecting features using a cross-attention mechanism. It is a light weight block which can be applied as a plug-in module for any feature fusion structure.
2) We implement FTB on YOLOX-s and YOLOX [4]. The experiments on COCO [5] and CrowdHuman [6] demonstrate its effectiveness.

II. Related Works

A. One-stage Object Detector

One-stage object detectors are classified as proposal-based methods that generate predictions from pre-defined anchors. It views the object detection as a regression problem which forms an end-to-end structure to get the prediction result all at once. The represented method are YOLO series [4], SSD [7]. These detectors typically consist of three parts: backbone, neck and head. The backbone serves as the feature extraction component, while the neck rearranges the feature maps generated by the backbone and sends them to the head. This simple design ensures low detection latency while maintaining considerable detection accuracy.

B. Multi-scale Detection

Multi-scale detection is a fundamental technique in computer vision, essential for capturing objects of varying sizes in an image. Traditionally, feature pyramids were created by extracting features from images with different input sizes. However, this approach resulted in a significant increase in inference time. Moreover, using a single-scale feature map often led to low recall and imprecise location predictions, limiting the object detectors' overall performance.

In modern approaches, dense prediction models leverage multi-scale features directly generated by the backbone feature hierarchy for multi-scale detection. For example, SSD utilized

979-8-3503-4017-4/23/$31.00 ©2023 IEEE

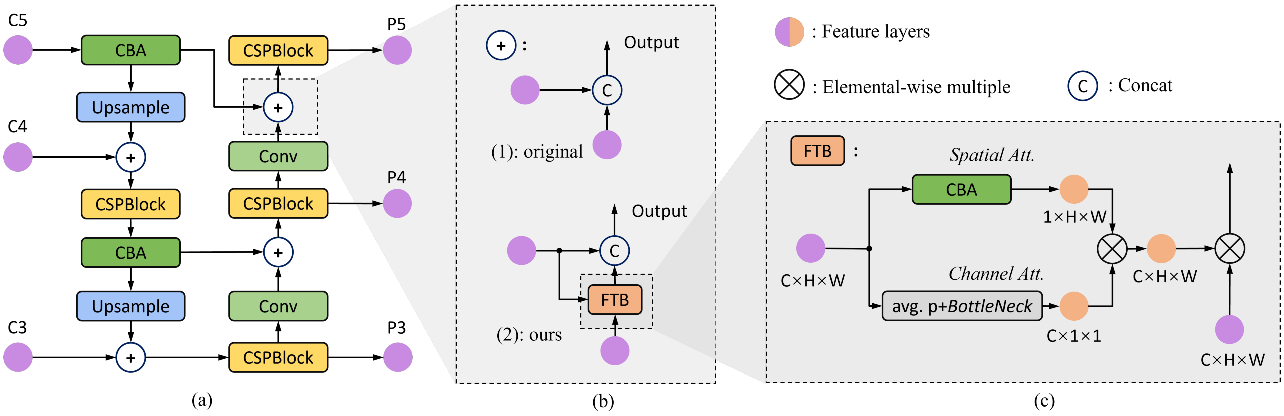

Fig. 1. (a) The overview of the PANet in YOLOX. CBA stands for the sequence of ConvBlock, BatchNorm(BN), and Activation function. It uses nearest-neighbor interpolation as the up-sampling method and adopts a simple convolution block with a stride of 2 for down-sampling. The detailed structure of CSPBlock can be found in [4]. (b) The upper (1) represents the original multi-scale feature fusion strategy. The lower (b) is the proposed FTB. (c) The illustration of FTB. The channel attention is first calculated by channel average pooling and then passed through a bottleneck structure that squeezes the channels from C to $C/16$ and then recovers to C.

multiple intermediate feature maps from the backbone. The FPN not only utilizes bottom-up features but also conducts top-down feature fusion, enabling stronger semantic information from higher pyramid levels to reinforce low-level features. This combination results in feature maps with both low-level details and abstract semantic information, leading to improved accuracy and robustness of the detector.

Comparing to FPN, the PANet incorporates additional bottom-up connections after the top-down stage, enabling more extensive feature fusion. PANet is often employed as a neck in general one-stage object detectors. As shown in Fig. 1(a), C_n and P_n represent the feature maps generated from the backbone and after the neck, respectively. Here, n indicates the down-sampling ratio of 2^n compared to the input shape. In typical YOLO-series detectors, the neck part uses C_3, C_4, C_5 as input and P_3, P_4, P_5 as the output.

C. Object Detection Benchmarks

COCO [5] is considered to be the most used benchmark for general object detection. The detector's score on the COCO is highly important and widely recognized as a standard for evaluating its performance.

CrowdHuman [6] is published for better evaluating detectors in crowd scenarios. With average number more than 22 persons per image, it is both a valuable pedestrian detection training material and a challenge benchmark for the object detectors.

III. Proposed Method

In this section, we propose a novel feature fusion approach called Feature Transfer Block (FTB) for multi-scale detection. Unlike traditional methods that implicitly select and merge feature maps, FTB adopts a cross-attention-like architecture to explicitly select and transfer semantic information between different scales. The FTB acts as a plugin module, seamlessly replacing the previous fusion components as shown in the Fig. 1(b).

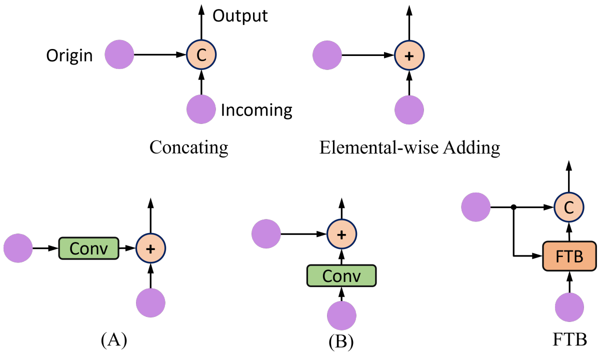

Fig. 2. Various architecture of feature fusion strategy. The upper two are the traditional feature fusion strategy. Type A: extra convolution layer for original layer. Type B: extra convolution layer for incoming layer. FTB: feature transfer block

To facilitate the understanding of the fusion process, we provide definitions for the involved components. When considering two input layers, one is designated as the ”incoming layer” if it is either up-sampled or down-sampled from different levels. The layer at the same level as the output is referred to as the ”original layer.”. The fusion process is to transfer the semantic information from the incoming layer to the original layer.

Feature fusion involves the selection and rearrangement of semantic information from different domains in the feature pyramid, where each domain represents a different scale. In complex scenarios, the prediction of a target benefits from not only its own texture but also global context information from different scales. Therefore, explicitly selecting and merging features before fusion leads to a promising approach.

To achieve the decoupling of feature extraction and feature fusion, we introduce additional modules: type A, type B, and the FTB (Fig. 2). These modules, which consist of simple convolution layers, significantly contribute to the effectiveness of our proposed approach. Type A and type B serve as naive implementations and contrast groups; they simply add additional layers to the original input and incoming input, respectively, with the hope of selecting semantic information. On the other hand, the FTB explicitly selects semantic information from the incoming layer with reference to the original layer, facilitating more effective and informative feature fusion.

The PANet for YOLOX, depicted in Fig. 1(a), comprises four fusion modules: $f_{5\rightarrow4}$, $f_{4\rightarrow3}$, $f_{3\rightarrow4}$, and $f_{4\rightarrow5}$. Each module corresponds to transferring features from level 5 to 4, level 4 to 3, level 3 to 4, and level 4 to 5, respectively. These fusion modules play a critical role in integrating multi-scale features, enhancing the overall detection performance.

FTB explicitly selects semantic information from incoming layer with the reference of original layer. We denote the original feature layer and the incoming feature layer as I_{gate} and I_{info}. Then the spatial attention $Att_{sp} \in \mathbb{R}^{1\times W\times H}$ and the channel attention $Att_{chn} \in \mathbb{R}^{C\times1\times1}$ can be calculated as

$$\begin{aligned} Att_{sp} &= Sigmoid(BN(Conv_{1\times1}(I_{gate}))), \\ Att_{chn} &= Sigmoid(BottleNeck(Avg.P(I_{gate}))). \end{aligned} \tag{1}$$

where BN represents the batch-norm layer. The channel attention is first calculated by channel average pooling and then passed through a bottleneck structure that squeezes the channels from C to $C/16$ and then recovers to C. Then the discriminate selection output is obtained by

$$I_{out} = (Att_{sp} \times Att_{chn}) \times I_{info}, \tag{2}$$

as shown in Fig. 1(c). This operation combines the spatial and channel attention masks (Att_{sp} and Att_{chn}) with the incoming feature layer (I_{info}) to produce the refined output feature map I_{out}.

IV. Experiments

A. Implement Details

During the training process on the COCO dataset, we strictly follow the training scheme outlined in [4]. Additionally, we utilize COCO's official tools for the model evaluation.

For the training on the CrowdHuman dataset, we keep most of the training settings but adjust the batch size from 64 to 32. Moreover, we adopt the evaluation metric used in [8] to assess the model's performance on the CrowdHuman dataset. MR^{-2} reflects the missing rate of detection per image and JI evaluates the degree of overlap between the predicted bounding box and the ground truth. A well performed detector is expected to have low MR^{-2} and high JI. Detailed definition can be found in [8].

B. Experiments Results

The first experiments is conducted on CrowdHuman. We start by training the YOLOX and YOLOX-s models to establish the baseline performance. We then proceed to evaluate the impact of our proposed Feature Transfer Block (FTB) on the performance of these models.

Table I presents a comprehensive comparison of the different fusion strategies and the inclusion of the FTB module for both YOLOX and YOLOX-s models on the CrowdHuman dataset. The baseline fusion strategy is the "Concat" method. We then evaluate two alternative fusion strategies, "Type A" and "Type B," which yield slightly improved results, showing AP values of 90.20% and 90.21%, respectively. The FTB module demonstrates its effectiveness by further boosting the performance, achieving an AP of 90.31%.

TABLE I
Comparsion experiments of different fusion strategy on CrowdHuman

Model	Fusion Strategy	AP ↑	MR^{-2} ↓	JI ↑
YOLOX-s	Concat(Baseline)	90.15	43.72	75.89
	Type A	90.20	43.56	75.79
	Type B	90.21	43.67	75.75
	FTB	90.31	43.47	75.73
YOLOX	Concat(Baseline)	92.19	39.88	77.80
	FTB	92.44	39.30	78.58

TABLE II
Comparsion experiments on COCO

Model	FTB	AP$_{val}$(%) ↑	AP$_{50}$	AP$_{75}$
YOLOX-s	×	40.4	59.3	43.7
	✓	41.0	59.7	44.2

Similarly, for the YOLOX model, the baseline fusion strategy leads to an AP of 92.19%. However, when incorporating the FTB module, we observe a significant overall performance improvement, with the AP increasing to 92.44%, MR^{-2} decreasing to 39.30%, and JI rising to 78.58%.

These results demonstrate that the inclusion of the FTB module enhances the performance of both YOLOX and YOLOX-s models on the CrowdHuman dataset, validating its effectiveness in facilitating feature transferability and improving object detection accuracy under extreme crowded targets scenario.

To demonstrate its improvement on general object detection, We then conduct the experiments of YOLO-s on the COCO.

Table II presents the comparison of the YOLOX-s model's performance with and without the FTB module. Under the baseline setting without the FTB module, our reproduced YOLOX-s model achieves its self claimed performance according to its official implementation [4]. However, the adaptation of the FTB module improves the model's AP by 0.6% which we consider to be significant on COCO benchmark.

These results indicate that the FTB module plays a beneficial role in general object detection, further reinforcing its effectiveness in feature transfer and multi-task learning scenarios.

C. Visualization Results

To provide insights into the effectiveness and capabilities of our proposed Feature Transfer Block (FTB), we utilize visualization techniques, specifically heatmap generation, to analyze the spatial attention within the FTB module. These

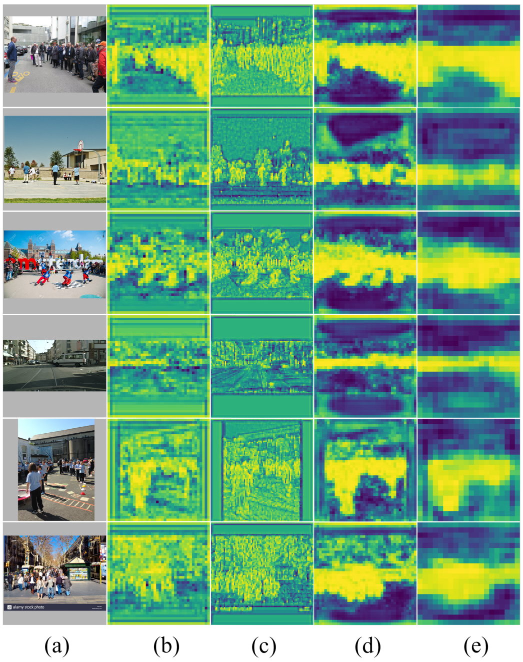

Fig. 3. Visualization of spatial attention within the Feature Transfer Block (FTB) on input images from the CrowdHuman Dataset. (a) is the input images in letter-box. The heatmaps (b), (c), (d), and (e) correspond to spatial attention visualization on $f_{5\to4}$, $f_{4\to3}$, $f_{3\to4}$, and $f_{4\to5}$, respectively.

heatmaps offer valuable visualizations of the attention mechanism's internal workings, showcasing the regions of interest and focus during the model's decision-making process.

We visualize the regions of high attention within the FTB when applied to the CrowdHuman dataset to demonstrate the efficacy of the FTB's spatial attention mechanism. Fig. 3 presents the visualization of spatial attention on feature transfers between different layers in YOLOX-s, including $f_{5\to4}$, $f_{4\to3}$, $f_{3\to4}$, and $f_{4\to5}$. These heatmaps indirectly validate the attention mechanism's ability to effectively capture and prioritize relevant features, leading to improved model predictions.

Furthermore, Fig. 4 showcases specific examples of hot maps generated from the spatial attention within the FTB on the feature transfer $f_{3\to4}$. Notably, the hot maps clearly indicate that the spatial attentions are focusing on crowded regions, illustrating the FTB's capability to capture important information in complex scenes.

These visualizations bolster the efficacy and interpretability of the spatial attention mechanism within the proposed FTB, offering compelling evidence for its contribution to improving the feature fusion process.

V. Conclusion

In this paper, in order to address the need for an improved multi-scale feature fusion strategy, we propose the Feature Transfer Block to enhance the performance of the yolo-like one-stage detector. Unlike traditional methods that implicitly select and merge feature maps, FTB adopts a cross-attention-like architecture to explicitly select and transfer semantic information between different pyramid levels. It is both computationally efficient and easy to integrate in place of the current fusion modules. Our experiments on the CrowdHuman datasets indicate that the implementation of the FTB can moderately improve the performance of an one-stage object detector in dense crowd scenario. The comparsion experiments on COCO also demonstrate FTB's robustness which contributes to the general object detection. What's more, the visualization result of the spatial attention in FTB clearly shows its effectiveness in focusing the important area.

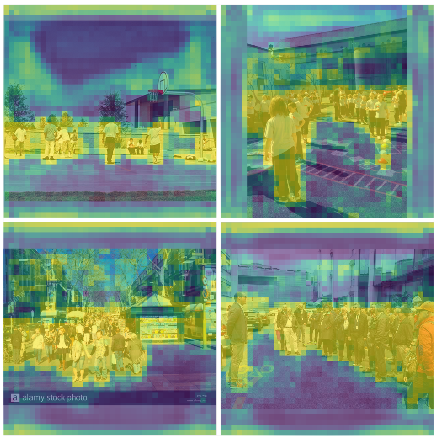

Fig. 4. Examples of hot maps generated from the spatial attention within the FTB on feature transfer $f_{3\to4}$. The hot maps clearly indicate the spatial attentions are focusing on the crowds, highlighting the FTB's ability to capture relevant features in complex scenes.

Overall, this paper proposes FTB as a novel feature fusion strategy, showcasing its moderate performance improvements on COCO and CrowdHuman benchmarks. These results validate the effectiveness of FTB in enhancing the overall performance of one-stage object detectors.

References

[1] T. Lin, P. Doll´ar, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature Pyramid Networks for Object Detection," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.936-944, Jun. 2017.

[2] S. Liu, L. Qi, H. Qin, J. Shi and J. Jia, "Path Aggregation Network for Instance Segmentation," in 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.8759-8768, Jun. 2018.

[3] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10778–10787, IEEE, Jun. 2020.

[4] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," arXiv preprint arXiv:2107.08430, Jul. 2021.

[5] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll ar, and C. L. Zitnick, "Microsoft coco: Common objects in context," in Computer Vision – ECCV 2014, pp. 740–755, May 2014.

[6] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun, "Crowdhuman: A benchmark for detecting human in a crowd,"arXiv preprint arXiv:1805.00123, Apr. 2018.

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, A. Berg, B. Leibe, J. Matas, N. Sebe, M. Welling, "SSD: Single Shot MultiBox Detector," in Computer Vision – ECCV 2016, pp.21-37, Sep. 2016

[8] X. Chu, A. Zheng, X. Zhang, and J. Sun, "Detection in Crowded Scenes: One Proposal, Multiple Predictions," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12211–12220, Jun. 2020.