Information Sciences 741 (2026) 123223

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A novel multi-granularity context-adaptive downsampling convolution for low-resolution images

Zejun Gu[a], Zhong-Qiu Zhao[a,*], Hao Shen[b], Zhao Zhang[a], De-Shuang Huang[c]

[a] School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, 230009, Anhui, China
[b] School of Public Security and Emergency Management, Anhui University of Science and Technology, Hefei, 231131, Anhui, China
[c] Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, 315201, Zhejiang, China

HIGHLIGHTS

- We identify key causes of CNN performance drops on low-resolution images.
- Detail loss and weak multi-granularity modeling hinder feature learning.
- We propose MCDC, a plug-and-play context-adaptive downsampling convolution.
- MCDC preserves information and models multi-granularity features effectively.
- Experiments on CIFAR-10, TinyImageNet, COCO, and MPII show consistent gains.

ARTICLE INFO

Keywords:
Low-resolution images
Multi-granularity
Context-adaptive
Downsampling convolution

ABSTRACT

In practical applications, low-resolution input images are common, yet existing models often degrade significantly in such scenarios because they fail to learn sufficient informative features. Most models rely on strided convolutions or pooling layers for downsampling, which further exacerbates information loss. To address this, we propose a Fine-grained Information Preservation (FIP) submodule to mitigate the loss of local details during downsampling. Furthermore, traditional convolutions usually focus only on local regions, making it difficult to model long-range contextual dependencies. To overcome this limitation, we design a Coarse-grained Context Aggregation (CCA) submodule to further exploit contextual information at multiple granularities. Considering that the importance of contextual information varies across scenarios, we further introduce a Context-Aware Granularity Adapter (CAGA), which dynamically adjusts the contributions of different granularities to enhance feature extraction. Distinct from prior multi-dilated or space-to-depth designs, our approach explicitly models adaptive multi-granularity fusion to flexibly fuse fine-grained details and coarse-grained contextual information. Therefore, we present a novel Multi-granularity Context-adaptive Downsampling Convolution (MCDC) as a replacement for traditional downsampling layers. Our approach requires no task-specific design and integrates seamlessly into various low-resolution models. Extensive experiments on benchmark datasets (TinyImageNet, COCO, MPII, and CIFAR-10) validate its effectiveness. Code is available at: https://github.com/guzejungithub/MCDC.

* Corresponding author.
Email address: z.zhao@hfut.edu.cn (Z.-Q. Zhao).

https://doi.org/10.1016/j.ins.2026.123223
Received 30 October 2025; Received in revised form 7 February 2026; Accepted 7 February 2026
Available online 12 February 2026

1. Introduction

In practical computer vision applications, low-resolution input images are frequently encountered [1]. First, the distance between the camera and the target object often varies, frequently resulting in distant objects that appear small and low-resolution. Second, due to economic constraints, limited technical conditions, and environmental factors, many imaging devices used in real-world applications are incapable of capturing high-resolution images. Third, considering transmission and storage costs, numerous images must be processed at reduced resolution. Finally, devices such as smartphones, autonomous driving systems, and IoT edge devices are limited by computational resources, making it challenging to handle high-resolution images. Therefore, effectively processing low-resolution inputs is a fundamental issue in practical engineering applications.

Despite the recent success of Transformer-based models, CNNs remain indispensable. CNNs excel at capturing local features such as textures and edges, are lightweight and efficient, and are better suited to small-scale datasets. Furthermore, CNNs benefit from a mature technological ecosystem, including abundant optimization techniques, pretrained models, and hardware acceleration (GPU/TPU). ***Consequently, CNNs continue to be the preferred choice for industrial and edge applications.***

However, current CNN models often suffer from significant performance degradation under low-resolution conditions. As shown in Fig. 1, these models perform poorly across different tasks when input images are low-resolution, failing to meet practical requirements. Therefore, improving CNN performance in low-resolution scenarios is critical.

Recently, some studies have attempted to alleviate performance degradation caused by low-resolution input images from various perspectives. Some approaches use knowledge distillation to transfer information from high-resolution teacher models to low-resolution student models [2], enhancing recognition performance. Several works introduce uncertainty modeling to improve performance on low-resolution inputs [3]. In addition, image super-resolution has been leveraged to enhance input features [4], while multi-scale feature enhancement strategies have been proposed to better represent low-resolution targets [5]. Moreover, some works focus on improving CNN architectures to enhance small-object detection [6]. Nevertheless, current approaches still face limitations: ***(1) Most methods are task-specific and lack generality. (2) Many approaches rely on multi-stage pipelines, repeated training, and extensive hyperparameter tuning, which are impractical in engineering applications. (3) The performance gains are often limited. (4) Some methods cannot handle both high- and low-resolution scenarios.*** To address all these challenges, we propose a Multi-granularity Context-adaptive Downsampling Convolution, named MCDC, which is ***generalizable, plug-and-play, and directly replaces traditional downsampling layers*** in CNNs, achieving significant performance improvements across multiple vision tasks.

MCDC consists of three submodules: a Fine-grained Information Preservation (FIP) submodule, a Coarse-grained Context Aggregation (CCA) submodule, and a Context-Aware Granularity Adapter (CAGA). For convenience, we adopt human pose estimation as the base task to demonstrate our method, due to its sensitivity to both fine-grained information (keypoints) and coarse-grained contextual information (human body structure).

First, the fine-grained information preservation submodule reduces information loss during downsampling. As illustrated in Fig. 2(a), when a convolution window of a kernel size of 2 (blue box) slides over the image with a stride of 3, some regions are skipped. In the left figure, the blue areas indicate the regions where the convolution is actually applied, while the uncovered areas are ignored, representing the lost image information. The information preserved during the convolution process is shown in the right figure. As shown in Fig. 2(b), max pooling only retains the maximum value within a region, while the other values and their distribution within the region are lost. Our proposed FIP submodule supplements the lost spatial information into channels while reducing spatial size, thereby ensuring effective feature learning during size transformation.

Second, the coarse-grained context aggregation submodule captures contextual information at multiple granularities. As shown in Fig. 3, the fine-grained submodule only captures local information and cannot model long-range dependencies. Inspired by the human visual system, which often relies on overall object contours rather than local details to recognize low-resolution blurred images, we adopt the CCA submodule that extracts contextual information with large receptive fields to capture long-range semantic relationships.

Finally, the context-aware granularity adapter fuses multi-granularity features by adaptively weighting them according to their importance. For human pose estimation, local keypoint information is more important in unoccluded cases, whereas coarse-grained body structure information is more critical when occluded. For image classification, foreground information is more important than

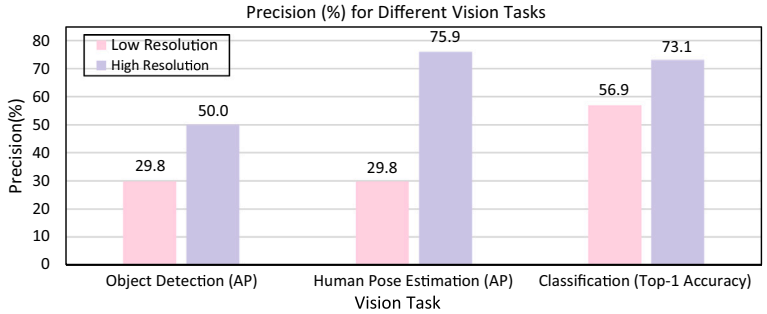

Fig. 1. Comparison of the performance of state-of-the-art (SOTA) models on high- and low-resolution images for various tasks. The performance of current models degrades significantly on low-resolution tasks.

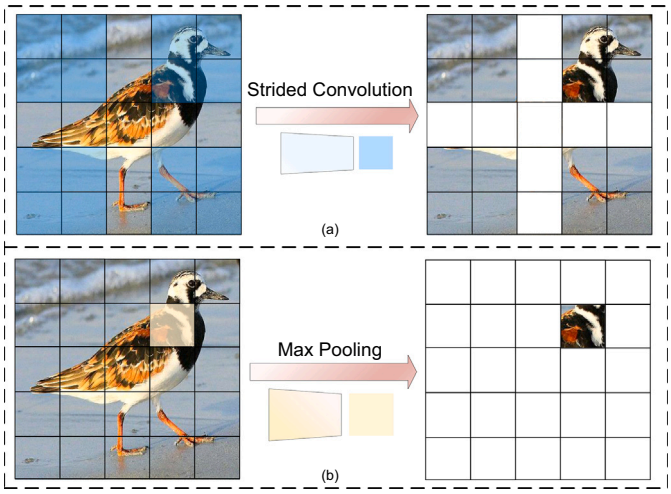

Fig. 2. Illustrative examples showing how traditional strided convolutions and pooling layers lead to image information loss. **(a)** Feature information extracted using a 2×2 convolution (blue box) with a stride of 3 for downsampling. **(b)** Feature information extracted using a max-pooling layer (yellow box).

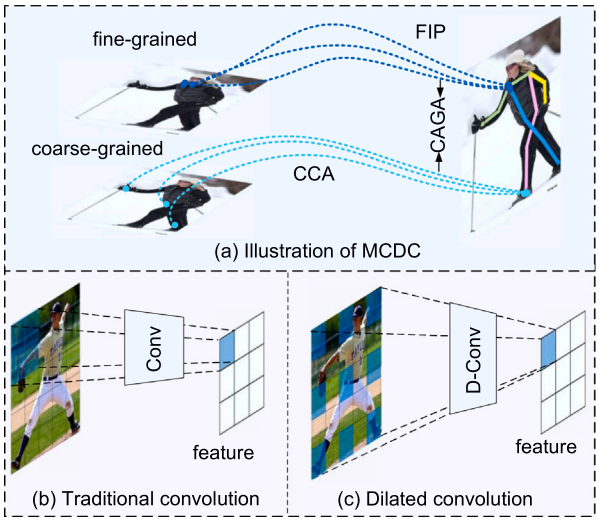

Fig. 3. (a) Illustrative examples of MCDC, showing the fusion of fine-grained keypoint visual cues and coarse-grained skeletal information. **(b)** Principle of traditional convolution. **(c)** Principle of dilated convolution.

background. Visualization examples are provided in the supplementary materials. The adapter dynamically adjusts the contributions of different granularities to fully exploit effective features.

The main contributions of this work are as follows:

- We provide a principled analysis of the performance degradation of CNNs under low-resolution settings, revealing a previously underexplored mechanism: conventional downsampling operations in LR images inherently alter the structural organization of visual information, leading to further loss of fine-grained detail information and insufficient modeling of multi-granularity representations, which fundamentally constrain effective feature learning.
- We reinterpret downsampling as a multi-granularity context modeling problem, rather than a purely resolution-reduction operation. Based on this formulation, we propose MCDC, a general context-adaptive downsampling convolution that explicitly preserves fine-grained information while enabling adaptive fusion across different semantic granularities, without relying on task-specific priors or additional training paradigms.
- Extensive experiments on COCO [7], MPII [8], TinyImageNet [9], and CIFAR-10 [10] demonstrate the robustness and generality of our method. It consistently improves performance across multiple tasks, datasets, input resolutions, and backbone architectures. Notably, it achieves significant gains of $+10.7$ AP in object detection and $+8.7$ AP in human pose estimation.

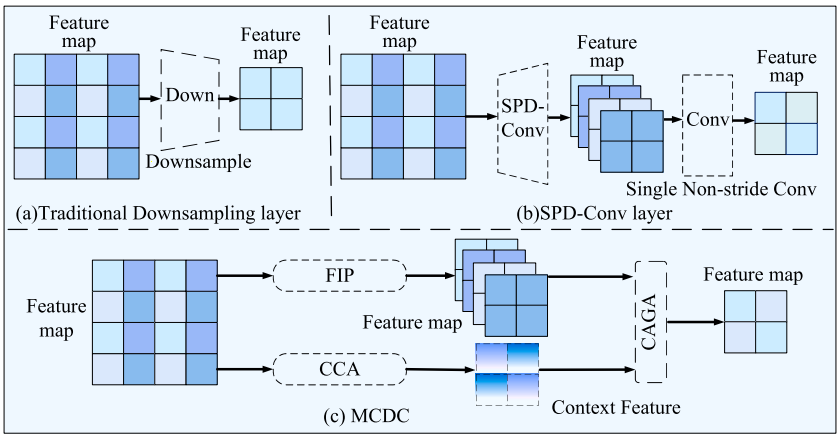

Fig. 4. Conceptual comparison: **(a)** conventional downsampling layer, **(b)** SPD-Conv layer [16], and **(c)** our proposed MCDC module. Each small rectangle represents the local information from its corresponding region.

2. Related work

2.1. Low resolution images

Low-resolution images are common in real-world applications, while existing models exhibit significant performance degradation when processing them. To address this issue, researchers have proposed several methods. Haris et al. [11] studied the role of image super-resolution (SR) in low-resolution object detection and proposed a task-driven end-to-end SR framework that incorporates detection loss into the training of the SR sub-network, thereby significantly improving the accuracy of object detectors on low-resolution images. Lo et al. [3] introduced probabilistic data uncertainty learning and an emotion wheel to address performance degradation in facial expression recognition under low resolution and label ambiguity, and further introduced a neutral expression constraint to enhance feature robustness. Zhang et al. [4] presented SuperYOLO, which fuses multimodal data and employs a super-resolution (SR) branch to improve detection of multiscale small objects in low-resolution remote sensing images. The SR branch is discarded during inference to reduce computational cost, achieving a favorable trade-off between accuracy and speed. Deng et al. [5] proposed EFPN, which adds a high-resolution pyramid level to the feature pyramid network to enhance small object detection. They further introduced a Feature Texture Transfer (FTT) module and a cross-resolution distillation mechanism to improve the perception of small object details, and designed a foreground-background balanced loss to alleviate the area imbalance issue. Li et al. [1] designed the TELOD framework, which employs a lightweight Task Disentanglement Enhancement Network (TDEN) for super-resolution reconstruction and integrates an Auxiliary Feature Enhancement Head (AFEH) with high-resolution priors. Shi et al. [6] proposed YOLO-KRM to address low detection accuracy for small targets and complex backgrounds in remote sensing images, enhancing feature representation and global context capture by incorporating the Kolmogorov–Arnold network, receptive field attention convolution, and a hybrid local channel attention module.

However, these works focus solely on specific tasks and do not propose a general module for low-resolution scenarios. In this work, we propose a universal convolution, MCDC, specifically designed for the characteristics of low-resolution images and applicable to various visual tasks. It is plug-and-play, allowing direct integration into CNN models to enhance their performance.

2.2. Novel convolution

Convolution is a fundamental operator in deep neural networks, primarily used for feature extraction and spatial downsampling. Its key parameters, such as kernel size and stride, respectively control the receptive field and the degree of resolution reduction. To meet diverse computational and architectural requirements, various convolution variants have been proposed. Manessi et al. [12] proposed WD-GCN and CD-GCN, which combine graph convolutional units with LSTMs for vertex and graph sequence classification on dynamic graphs. Cao et al. [13] proposed DO-Conv, a depthwise over-parameterized convolutional layer that augments conventional convolutions with depthwise kernels to improve CNN performance across vision tasks without increasing inference complexity. Xie et al. [14] presented HCN-WDI, a hypergraph convolutional network for wafer defect identification that models hyper-relations among samples, incorporates data augmentation, and leverages conventional image classifiers as feature extractors to achieve high classification accuracy. Chen et al. [15] introduced a residual estimation-oriented dynamic spatio-temporal graph convolutional network (DSTGCN) for wind speed interval prediction, which models time-varying spatial correlations and extracts spatio-temporal features to provide accurate probabilistic forecasts while addressing quantile crossing issues.

However, none of these works have designed convolutions specifically for low-resolution visual tasks. Sunkara et al. [16] proposed SPD-Conv to enhance performance in such scenarios. However, it suffers from several limitations (see Fig. 4b): (1) It performs dimensional transformation only through a single non-strided convolution, resulting in insufficient exploitation of fine-grained information. (2) It does not explicitly model contextual structures, nor does it incorporate an adaptive mechanism for fusing multi-granularity information. In contrast, we revisit the downsampling process in low-resolution scenarios from a modeling perspective and propose a

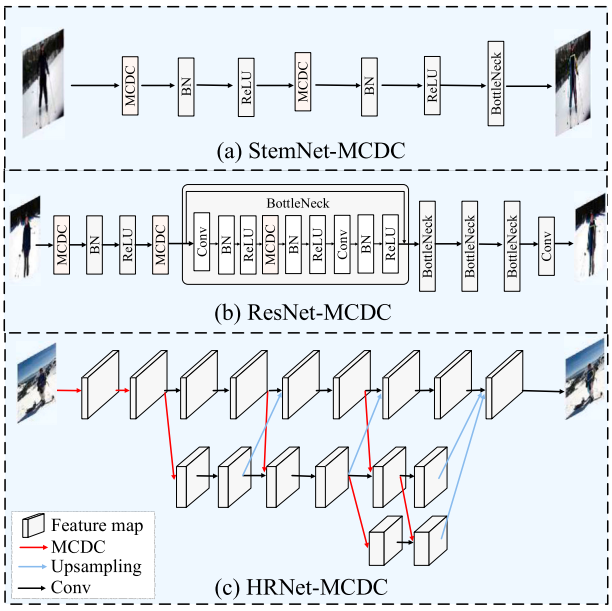

Fig. 5. The overall architecture of **(a)** StemNet-MCDC, **(b)** ResNet-MCDC, and **(c)** HRNet-MCDC, where MCDC replaces the standard downsampling layer in each backbone.

multi-granularity context-adaptive downsampling convolution (MCDC). It fully leverages fine-grained information through the FIP submodule, and adaptively utilizes multi-granularity information through the CCA submodule and the CAGA submodule (see Fig. 4c).

3. Methods

We first introduce the overall framework of our MCDC module, and then provide a detailed description of the proposed module.

3.1. Overall framework

Fig. 4(c) provides a framework for the proposed module. The module is composed of a Fine-grained Information Preservation (FIP) submodule, a Coarse-grained Context Aggregation (CCA) submodule, and a Context-Aware Granularity Adapter (CAGA). The FIP submodule includes a Space–Channel Transformation (SCT) unit and a Deep Information Extraction (DIE) block. It is introduced in Section.3.2. The CCA submodule consists of a Multi-Receptive Information Extraction (MRIE) unit and a Deep Information Extraction (DIE) block. We describe it in Section. 3.3. CAGA is described in Section. 3.4.

The module we propose can replace the traditional downsampling layer in the current models, thereby enabling these models to achieve better performance in low-resolution tasks. Fig. 5 illustrates the overall architecture of applying the MCDC module to StemNet [17], ResNet [18], and HRNet [19].

3.2. Fine-grained information preservation (FIP) submodule

For low-resolution visual tasks, existing CNN models commonly use downsampling layers to reduce redundant information and alleviate computational costs. The downsampling layer results in the loss of much fine-grained information in low-resolution images. To address this problem, we design a Fine-grained Information Preservation (FIP) submodule.

The operation of traditional downsampling layers can be expressed as:

$$X' = Down(X), \tag{1}$$

where $X \in \mathbb{R}^{H\times W\times C}$ denotes the input feature map of the downsampling layer, and H, W, and C are the height, width, and the number of channels of the input feature, respectively; $X' \in \mathbb{R}^{H'\times W'\times C'}$ denotes the output feature of the downsampling layer, and H', W', C' are the height, width, and the number of channels of the output feature maps, respectively; $H' = H/2$, $W' = W/2$, $C' = C$. To enable direct replacement of downsampling layers, MCDC is required to maintain consistency with the replaced downsampling layers in terms of both input and output feature sizes. Thereby, we design a space-channel transformation (SCT) unit. As depicted in Fig. 6(b), for any input feature $X \in \mathbb{R}^{H\times W\times C}$ of a downsampling layer, we sample the spatial domain at a fixed interval of two pixels:

$$X' = Sample(X), \tag{2}$$

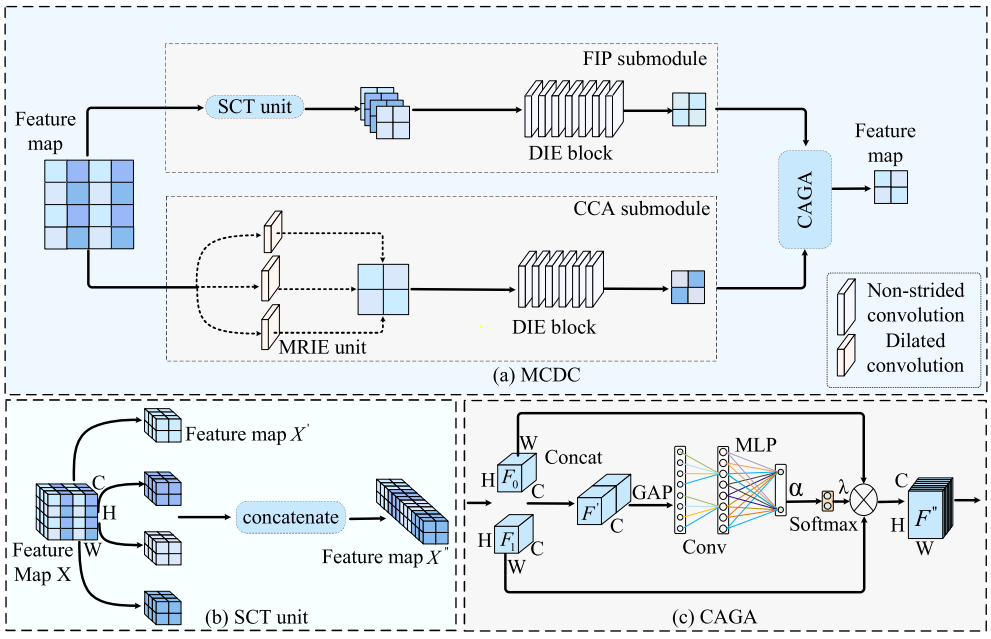

Fig. 6. (a) Overview of our proposed MCDC module. **(b)** Illustration of SCT unit. **(c)** Context-aware granularity adapter (CAGA).

where $Sample(\cdot)$ denotes an interval sampling operation. It can be concretely expressed as:

$$x_1' = \begin{pmatrix} x_{0,0} & x_{0,2} & \cdots \\ x_{2,0} & \ddots & \vdots \\ \vdots & \cdots & x_{w-2,h-2} \end{pmatrix}, \tag{3}$$

$$x_2' = \begin{pmatrix} x_{1,0} & x_{1,2} & \cdots \\ x_{3,0} & \ddots & \vdots \\ \vdots & \cdots & x_{w-1,h-2} \end{pmatrix}, \tag{4}$$

$$x_3' = \begin{pmatrix} x_{0,1} & x_{0,3} & \cdots \\ x_{2,1} & \ddots & \vdots \\ \vdots & \cdots & x_{w-2,h-1} \end{pmatrix}, \tag{5}$$

$$x_4' = \begin{pmatrix} x_{1,1} & x_{1,3} & \cdots \\ x_{3,1} & \ddots & \vdots \\ \vdots & \cdots & x_{w-1,h-1} \end{pmatrix}, \tag{6}$$

where x_i' $(i = 1,2,3,4)$ is the sampling result with different coordinate points as starting points, h and w represent the height and width of the input feature map, respectively. Then we concatenate x_i' along the channel dimension:

$$X'' = Concat[x_1', x_2', x_3', x_4'], \tag{7}$$

where $Concat[\cdot]$ is the concatenation operator. This unit achieves feature size reduction without compromising feature information.

Subsequently, we further employ a Deep Information Extraction (DIE) block, which consists of multiple non-strided convolutions. This module fully learns feature representations while performing feature dimensional transformations to maintain channel consistency with the modified backbone. By combining the SCT unit and the DIE module, FIP fully exploits fine-grained feature information. Table. 5 provides strong evidence of the performance of the FIP submodule.

3.3. Coarse-grained context aggregation (CCA) submodule

Conventional convolution operations primarily focus on local fine-grained information. However, coarse-grained contextual information is also important. Taking pose estimation as an example, the skeletal structure information, the symmetrical relationship between joints, and the background information all contribute to inferring joint positions. This work designs a Coarse-grained Context Aggregation (CCA) submodule to capture coarse-grained information.

Dilated convolution enlarges the receptive field by introducing different dilation rates, enabling the extraction of multi-scale contextual features within a single layer. Based on this property, we construct a Multi-Receptive Information Extraction (MRIE) unit

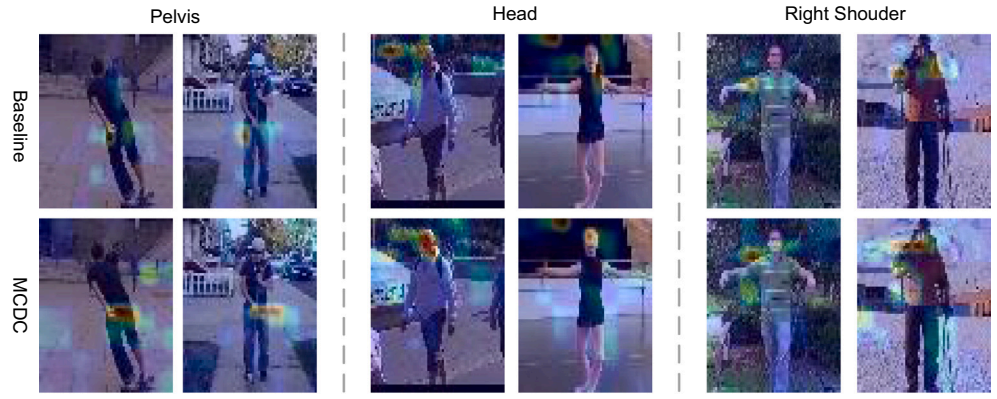

Fig. 7. Heatmap comparison between our method and standard convolution. The "pelvis", "head", and "right shoulder" indicate the target keypoints for pose estimation. The color intensity reflects the level of attention.

composed of dilated convolutions with multiple dilation rates to facilitate the extraction and interaction of coarse-grained contextual information.

Similar to the FIP submodule, to further exploit feature information, we also employ a DIE block within the CCA submodule. As shown in Fig. 6, the FIP submodule ensures the size transmission of information and fully learns feature information, while the CCA submodule further learns coarse-grained information. The combination of the two effectively addresses the issue of severe information deficiency in low-resolution images.

3.4. Context-aware granularity adapter (CAGA)

Typically, features of different granularities are fused via simple element-wise addition. However, their contributions vary. For example, in pose estimation, the local features of occluded keypoints mainly reflect the occluding objects rather than the keypoints themselves, and thus provide limited useful information. In this context, the structural information of the human body is more important and should be assigned higher weights during fusion.

To solve this problem, inspired by the Efficient Channel Attention (ECA) module [20], we propose a Context-Aware Granularity Adapter (CAGA). Fig. 6(c) illustrates CAGA. The inputs to CAGA are the output feature maps of different granularities, and CAGA assigns different weights to each granularity based on the corresponding feature information.

As shown in Fig. 6(c), initially, concatenate the two feature maps along the channel dimension and perform global average pooling to obtain 1D contextual features of:

$$F' = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} [F_0, F_1](i, j), \tag{8}$$

where F_0, F_1 are the output feature maps of each granularity respectively, $[\cdot]$ indicates the concatenate operation of feature maps along the channel dimension, and H, W are the height and width of the feature map respectively. The global pooling operations are advantageous for capturing spatial global information of features, which thus aids in the fusion of different features. Then F' is fed into a one-dimensional convolution and then into a fully connected layer to generate the weight matrix:

$$\alpha = MLP(Conv(F')), \tag{9}$$

where $Conv$ is a 1x1 convolution with an adaptive number of output channels. $MLP(\cdot)$ consists of a ReLU function and an FC layer with 2 output channels. The convolution operation can fully learn the correlation between the feature channels. The MLP serves two purposes: firstly, it reduces the number of feature channels to match the number of granularities. Secondly, it complements convolution operations, further enhancing the learning of channel correlations and thereby optimizing weight allocation. Finally, a Softmax operation is applied to the weight matrix to obtain the weight coefficients for different granularities:

$$\lambda = Softmax(\alpha), \tag{10}$$

where λ is a sequence of two-dimensional weight coefficients. The features of different granularities are combined with their corresponding weights to obtain the final fusion result:

$$F'' = \lambda_0 \cdot F_0 + \lambda_1 \cdot F_1, \tag{11}$$

where λ_0 and λ_1 are the coefficients corresponding to each granularity, respectively, and F'' is the fused features. F'' is then fed into the subsequent network structure. CAGA can learn spatial global information and channel correlations from each granularity, adaptively assigning weights to them, and ultimately producing a more effective output.

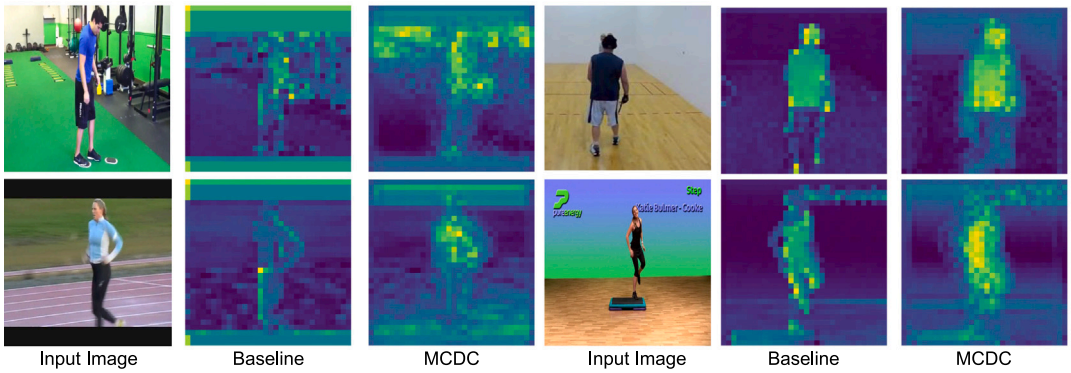

Fig. 8. Comparison of feature visualizations obtained at the first downsampling stage between our method and the baseline.

4. Analysis

4.1. Grad-CAM heatmap analysis

Grad-CAM [21] generates heatmaps of convolutional neural networks by computing the gradient information of the target class, providing a visual explanation of the model's decision-making process. It can intuitively reveal the key regions that the model relies on when making predictions. We divide the coordinate space into several bins, with each bin corresponding to a class. This transforms the keypoint prediction problem into a classification problem, allowing Grad-CAM to be applied to human pose estimation. We visualized heatmaps for different target keypoints on the last convolutional layer. The reason for visualizing the last convolutional layer is that its results directly reflect the basis of the model's predictions. As shown in Fig. 7, for the baseline method, the attention is not concentrated on the target region—it is either inaccurately localized or insufficiently concentrated. In contrast, after applying our proposed MCDC, the attention is densely concentrated on the target regions. This indicates that MCDC can retain more useful information, thereby guiding the model to focus on important regions and enhancing the interpretability of keypoint predictions.

4.2. Downsampling analysis

We visualized the feature activation maps obtained from the input images after the first downsampling. Specifically, the downsampled feature maps were averaged along the channel dimension to generate the corresponding feature activation maps. In the feature activation maps, the brightness of each pixel represents the overall activation intensity at that spatial location, *i.e.,* the combined response of all convolutional kernels at that position. It contains multi-dimensional effective information such as textures, contours, corners, and edges, and reflects the spatial distribution of feature strength. The higher the brightness at a given point, the more sufficient the effective information. For subsequent processing in the model, richer effective feature information facilitates more accurate predictions, especially in low-resolution tasks.

As shown in Fig. 8, our proposed method differs from traditional downsampling convolution in the following ways: for foreground regions, the brightness in the feature activation maps generated by our method is higher, indicating that our method preserves more effective feature information. In addition, traditional downsampling layers pay less attention to background regions, resulting in lower brightness in those areas. In contrast, the background regions in our method still exhibit relatively high brightness, demonstrating that our approach utilizes contextual information more effectively. Therefore, these visualizations provide an intuitive demonstration of the effectiveness and interpretability of our downsampling method.

4.3. Theoretical basis of the FIP submodule

This subsection presents a theoretical analysis of the FIP submodule. We first categorize downsampling layers according to their types and the relationship between the convolutional kernel size and the stride. Based on this categorization, we summarize two corresponding types of information degradation: direct information loss and underrepresentation of information. Finally, we conduct an in-depth analysis of these two types of information degradation.

Direct information loss. Pooling layers inevitably cause a direct loss of a significant amount of feature information. For example, max pooling discards all information except the maximum value, while average pooling eliminates the distribution characteristics of the features. Similarly, when the convolution stride exceeds the kernel size, information is directly lost during downsampling. In such cases, our method prevents direct information loss and ensures that effective features are fully utilized in subsequent processing.

Underrepresentation of information. When the convolution stride is less than or equal to the kernel size, downsampling does not directly result in information loss. However, single convolutions suffer from a limited receptive field and constrained feature representation capability. This limitation is particularly pronounced in low-resolution tasks, where the scarcity of fine-grained information further exacerbates the problem. According to related studies, multiple convolutions can enlarge the receptive field [22], enhance the network's non-linear representation capacity [18], and capture complex semantic features [23]. Therefore, our method enables more comprehensive exploitation of useful feature information.

Formal derivation of information losslessness. SCT enables lossless information transformation, and this subsection provides a theoretical proof. Given a feature X, and a transformation $f(\cdot)$, the transformation process can be expressed as:

$$Y = f(X). \tag{12}$$

According to [24], ***a transformation is information-lossless if and only if it is logically reversible, i.e., there exists a unique inverse transformation*** $f^{-1}(\cdot)$ ***such that***

$$X = f^{-1}(Y). \tag{13}$$

For the SCT unit, the inverse transformation can be explicitly constructed. Specifically, according to Eq. (7), the feature dimension of Y is rearranged while reducing the spatial resolution by a factor of four, without discarding any information. Following Eqs. (3)–(6), the features are interleaved and concatenated along the height and width dimensions:

$$X = \mathcal{I}(Y), \tag{14}$$

where $\mathcal{I}(\cdot)$ denotes a bijective interleaved concatenation operation. This operation directly corresponds to the inverse transformation $f^{-1}(\cdot)$ in Eq. (13). As a result, the original input X can be fully recovered. Therefore, SCT is a reversible transformation and is information-lossless. In contrast, for traditional strided convolution and pooling operations, the mapping is not injective, as there exist multiple inputs X that correspond to the same output $Y = f(X)$, making unique recovery impossible.

5. Experiments

The proposed MCDC is first evaluated on three representative computer vision tasks: human pose estimation, object detection, and image classification. To gain a deeper understanding of its internal mechanisms, we conduct ablation studies on the main components of MCDC and verify its compatibility using normal-resolution inputs. Subsequently, we perform visualization analyses to illustrate how MCDC makes predictions across different scenarios. ***Finally, we provide additional experimental results in the supplementary materials to demonstrate the generalization of MCDC under various settings.***

5.1. Datasets

COCO. The COCO dataset [7] is among the largest and most challenging benchmarks for both object detection and human pose estimation. For human pose estimation, it contains over 200K images and 250K human instances, each annotated with 17 keypoints. The dataset is divided into three subsets: 57K images for training, 5K images for validation, and 20K images for the test–dev set. For object detection, COCO provides the train2017 set with 118K images for training, val2017 with 5K images for validation, and test2017 with 40K images for testing.

MPII. The MPII dataset [8] comprises approximately 25K images containing over 40K annotated human instances, 29K of which are used for training and 11K for testing. The images were collected based on a well-established taxonomy of everyday human activities from YouTube videos. Following the common setup, we adopt the same train/validation/test split. Each person instance in the MPII dataset is annotated with 16 body joints.

Tiny ImageNet. The Tiny ImageNet [9] is derived from the ILSVRC-2012 classification dataset, which contains 200 classes. Each class comprises 500 training images, 50 validation images, and 50 test images, with each image sized $64 \times 64 \times 3$ pixels.

CIFAR-10. The CIFAR-10 [10] consists of 60K RGB images with a resolution of 32×32, including 50K training images and 10K test images. The dataset contains 10 classes, each with 6K images.

5.2. Evaluation metrics

Human pose estimation. For the COCO dataset, we adopt the standard Average Precision (AP) as the evaluation metric, which is calculated based on the Object Keypoint Similarity (OKS):

$$OKS = \frac{\sum_i exp(-d_i^2/2s^2j_i^2)\sigma(v_i > 0)}{\sum_i \sigma(v_i > 0)} \tag{15}$$

where d_i denotes the Euclidean distance between the i-th predicted keypoint and its corresponding ground-truth location, j_i represents a per-keypoint constant, v_i denotes the visibility flag, σ indicates the indicator function, and s is the object scale. We report standard average precision [9] and recall scores: AP^{50} (average precision at OKS = 0.50), AP^{75} (average precision at OKS = 0.75), AP (the mean of average precision scores at 10 positions, OKS = 0.50, 0.55,…, 0.90, 0.95), AP^M for medium objects, AP^L for large objects, and AR (average recall) at OKS = 0.50, 0.55, …, 0.90, 0.95.

For the MPII dataset, we use the standard Percentage of Correct Keypoints (PCK) metric, which quantitatively measures the proportion of predicted keypoints that fall within a normalized distance of the corresponding ground truth annotations. Specifically, PCKh@0.5 employs a threshold of 50% of the head diameter, effectively accounting for variations in individual body size, pose, and overall image scale.

Image classification. For the CIFAR-10 [10] and TinyImageNet [9], we adopt top-1 accuracy as the metric to evaluate classification performance.

Object detection. For accurate evaluation, we report the standard Average Precision (AP) on val2017, measured across IoU thresholds from 0.5 to 0.95 and for objects of different sizes (small, medium, large).

Table 1
Comparison with SOTA methods on the COCO validation dataset.

Method	Input size	Backbone	AP↑	AP^{50} ↑	AP^{75} ↑	AP^{M} ↑	AP^{L} ↑	AR↑
HRNet [19]	32 × 32	HRNet-W32	8.2	36.9	1.3	9.2	6.8	15.0
TokenPose [27]		HRNet-W32	14.0	48.2	3.4	15.2	12.5	21.7
CAL [28]		HRNet-W32	26.4	61.9	18.2	27.1	25.7	33.9
Rle [29]		HRNet-W32	24.4	–	–	–	–	–
Dark [30]		HRNet-W32	12.5	45.2	2.5	13.8	11.1	20.3
SimBase [25]		ResNet-152	4.4	21.1	1.0	5.3	3.2	9.0
PCT [31]		Swin-Base	1.3	4.6	10.0	1.0	1.2	3.1
Distillpose [32]		HRNet-W32	9.5	32.6	2.4	10.0	9.5	21.2
CAL [28]		HRNet-W48	29.1	65.5	21.4	29.9	28.6	36.7
SDPose [33]		Stem	4.4	22.3	1.0	5.2	4.2	12.0
DynPose [34]		Router	9.2	41.0	6.0	11.3	8.4	21.2
GTPT [35]		ShuffleNet	21.0	53.0	12.3	21.7	22.0	30.2
CDKD [36]		HRNet-W32	30.7	66.4	23.5	30.9	30.7	37.3
SimCC(*baseline*) [26]		ResNet-50	15.7	43.5	8.3	16.5	14.8	21.4
SimCC-MCDC(*ours*)		ResNet-50	**24.4**$_{+8.7}$	**58.1**	**16.8**	**25.3**	**23.5**	**30.1**
SimCC(*baseline*) [26]		HRNet-W32	29.8	65.6	22.5	30.0	29.9	36.3
SimCC-MCDC(*ours*)		HRNet-W32	**37.5**$_{+7.7}$	**73.9**	**34.2**	**37.5**	**37.9**	**43.7**
SimCC(*baseline*) [26]		HRNet-W48	30.6	65.8	24.2	31.4	29.9	37.2
SimCC-MCDC(*ours*)		HRNet-W48	**35.8**$_{+5.2}$	**70.7**	**32.2**	**36.8**	**35.1**	**42.0**
SimBase [25]	64 × 64	ResNet-152	30.3	67.6	22.6	30.6	30.5	36.2
Distillpose [32]		HRNet-W32	31.7	66.8	26.6	32.3	31.8	44.5
CAL [28]		HRNet-W32	58.5	87.0	64.8	57.7	59.9	63.6
PCT [31]		Swin-Base	11.8	37.4	4.80	12.2	12.0	17.8
Rle[29]		HRNet-W32	52.5	–	–	–	–	–
HRNet [19]		HRNet-W48	46.9	83.7	49.2	46.6	47.5	52.6
Tokenpose [27]		HRNet-W48	50.3	82.7	54.4	49.9	51.4	55.6
CAL [28]		HRNet-W48	60.6	88.1	68.4	59.5	62.3	65.5
Dark [30]		HRNet-W48	57.2	86.8	63.5	55.9	59.2	62.2
Tokenpose [27]		Stem	51.5	83.6	55.8	50.7	53.1	56.6
SDPose [33]		Stem	37.8	74.2	35.0	37.8	40.2	46.9
DynPose [34]		Router	45.4	81.9	46.6	45.7	47.7	55.0
GTPT [35]		ShuffleNet	40.2	73.2	39.7	39.6	43.3	48.7
CDKD [36]		HRNet-W48	61.1	86.9	68.4	59.9	62.9	65.6
RTMPose [37]		CSPNeXt-t	26.8	62.7	19.0	28.3	25.4	31.7
SimCC(*baseline*) [26]		ResNet-50	40.3	75.2	38.4	40.4	40.5	45.8
SimCC-MCDC(*ours*)		ResNet-50	**48.8**$_{+8.5}$	**80.4**	**51.7**	**48.6**	**49.7**	**53.7**
SimCC(*baseline*) [26]		HRNet-W32	56.4	85.7	62.8	55.6	57.9	61.3
SimCC-MCDC(*ours*)		HRNet-W32	**59.4**$_{+3.0}$	**86.0**	**66.2**	**58.6**	**61.1**	**63.9**
SimCC(*baseline*) [26]		HRNet-W48	58.6	85.9	64.9	57.8	60.5	63.4
SimCC-MCDC(*ours*)		HRNet-W48	**63.1**$_{+4.5}$	**88.0**	**70.7**	**61.6**	**65.4**	**67.5**

5.3. Implementation details

For all comparative methods, we replicate them following their original settings in the corresponding papers and code. We only modify the image input resolution and replace the downsampling layers with MCDC, while keeping all other settings unchanged. ***For all methods, we train and test under the same low-resolution inputs.*** All experiments are implemented using the PyTorch library on two NVIDIA GeForce 3080 Ti GPUs.

For human pose estimation, we adopt the top-down estimation pipeline. We use a commonly used person detector provided by SimpleBaselines [25] with 56.4 AP for the COCO val dataset. Since SimCC [26] has achieved the best performance under low-resolution input conditions in the latest research methods, we use SimCC as the baseline model. The training settings for our method, including the optimizer, learning rate, and data augmentation, follow those of SimCC. Specifically, we use the Adam optimizer with a base learning rate initialized at 1e-3, which is subsequently reduced to 1e-4 and 1e-5 at the 170th and 200th epochs, respectively. The batch size is set to 32. The data augmentation strategy is the same as that used in SimCC.

For image classification, we set the training hyperparameters on Tiny ImageNet as follows: SGD optimizer, learning rate of 0.01793, momentum of 0.9447, mini-batch size of 256, weight decay of 0.002113, and a total of 200 training epochs. Similarly, on the CIFAR-10 dataset, we train the model using the stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.1, a momentum of 0.9, a batch size of 128, and a weight decay of 0.0001. The training is conducted for a total of 200 epochs to ensure proper convergence and achieve overall optimal performance. For all image classification models, we use the same decay function as in ResNet [18] to gradually decrease the learning rate as the number of training epochs increase.

For object detection, we use the SGD optimizer with a momentum of 0.937 and a weight decay of 0.0005. During the first three warm-up epochs, the learning rate linearly rises from 0.0033 to 0.01, and then gradually decreases to 0.001 using a cosine decay strategy to ensure smooth training. Throughout the training process, the YOLOv5-based models use a batch size of 128.

Table 2
Comparison with SOTA methods on the MPII val dataset.

Method	Input Size	Backbone	Head.↑	Sho.↑	Elb.↑	Wri.↑	Hip.↑	Knee.↑	Ank.↑	PCKh@0.5↑
SimBase [25]	64 × 64	ResNet-152	89.905	85.258	73.854	64.058	76.216	68.628	62.659	75.353
OKDHP [38]		4-Stack HG	85.573	80.757	66.099	54.328	71.767	61.032	54.393	69.037
HRNet [19]		HRNet-W32	90.075	85.785	72.950	63.511	75.143	67.257	61.880	74.843
Dark [30]		HRNet-W32	89.802	87.568	75.251	64.112	78.726	69.616	63.628	76.612
CAL [28]		HRNet-W32	92.497	88.689	75.763	65.019	80.076	70.159	65.422	77.692
Tokenpose [27]		HRNet-W32	89.018	86.770	71.570	59.963	77.237	65.947	59.139	73.997
PRTR [39]		HRNet-W32	89.768	83.967	68.144	53.061	76.112	62.039	53.660	70.742
HRNet [19]		HRNet-W48	90.280	86.158	74.740	65.842	75.870	70.481	63.604	76.263
Dark [30]		HRNet-W48	90.314	88.604	76.785	65.976	80.699	73.040	66.840	78.426
Tokenpose [27]		HRNet-W48	89.427	87.704	74.093	62.072	78.311	67.036	59.896	75.298
SimCC(*baseline*) [26]		ResNet-50	88.165	81.165	63.082	48.210	70.279	57.265	51.771	66.903
SimCC-MCDC(*ours*)		ResNet-50	**90.859**	**85.224**	**68.638**	**54.945**	**74.243**	**61.757**	**56.377**	**71.410** $_{+4.507}$
SimCC(*baseline*) [26]		HRNet-W32	92.701	88.060	76.376	66.198	77.774	68.425	64.076	77.226
SimCC-MCDC(*ours*)		HRNet-W32	**92.838**	**89.623**	**79.461**	**70.174**	**80.336**	**72.033**	**67.524**	**79.758** $_{+2.532}$
SimCC(*baseline*) [26]		HRNet-W48	93.349	89.029	78.200	68.308	80.007	71.428	65.517	78.860
SimCC-MCDC(*ours*)		HRNet-W48	**94.134**	**89.589**	**78.728**	**69.848**	**80.301**	**71.388**	**67.690**	**79.617** $_{+0.757}$
SimBase [25]	32 × 32	ResNet-152	40.894	48.217	33.953	21.742	44.608	31.353	22.650	36.786
HRNet [19]		HRNet-W32	46.044	52.378	40.191	28.407	48.780	37.176	27.538	42.217
Dark [30]		HRNet-W32	38.984	61.804	46.191	31.867	61.277	45.234	28.839	48.137
CAL [28]		HRNet-W32	77.115	68.631	48.185	33.066	63.164	46.264	40.552	55.353
Tokenpose [27]		HRNet-W32	38.950	61.702	47.537	31.902	61.676	45.476	28.600	48.421
PRTR [39]		HRNet-W32	0.341	1.512	8.062	9.082	15.977	1.633	0.213	5.618
HRNet [19]		HRNet-W48	45.975	53.804	40.720	30.378	49.264	37.981	29.664	42.789
Dark [30]		HRNet-W48	39.666	64.878	49.497	33.168	64.445	48.116	30.042	50.370
Tokenpose [27]		HRNet-W48	40.177	64.640	49.225	33.873	62.576	46.222	28.932	49.779
SimCC(*baseline*) [26]		ResNet-50	71.317	64.096	44.111	26.554	57.088	39.593	33.561	49.794
SimCC-MCDC(*ours*)		ResNet-50	**80.082**	**73.064**	**50.844**	**30.784**	**63.753**	**47.089**	**38.521**	**56.406** $_{+6.612}$
SimCC(*baseline*) [26]		HRNet-W32	81.105	72.690	54.219	37.898	63.822	51.380	46.552	59.560
SimCC-MCDC(*ours*)		HRNet-W32	**87.074**	**79.721**	**61.633**	**47.921**	**69.275**	**56.297**	**51.015**	**65.948** $_{+6.388}$
SimCC(*baseline*) [26]		HRNet-W48	82.401	73.964	55.275	38.822	63.787	51.884	46.905	60.289
SimCC-MCDC(*ours*)		HRNet-W48	**86.937**	**80.146**	**62.468**	**48.570**	**69.206**	**56.076**	**52.031**	**66.282** $_{+5.993}$

5.4. Experimental results

5.4.1. Human pose estimation

COCO. We replicate existing state-of-the-art (SOTA) models under various low-resolution settings and evaluate our method on the COCO validation dataset. The experimental results are summarized in Table. 1. SimCC-MCDC achieves superior performance among all compared methods. For example, SimCC-MCDC with HRNet-W32 outperforms the previous dominant methods including HRNet [19], SimCC [26], SDPose [33] and TokenPose [27] at a 32 × 32 resolution. Similarly, our method with HRNet-W48 also achieves better results than the SimCC [26] method at 64 × 64 resolution. In different resolutions and backbones, our method achieves a consistent performance improvement. Notably, we surpass the baseline method (SimCC [26]) by 8.7 AP when the backbone is ResNet-50 and the input resolution is 32 × 32. It fully demonstrates that our proposed MCDC is an effective and versatile module for low-resolution human pose estimation.

MPII. The results on the MPII validation set are shown in Table. 2. We also conduct experiments at different resolutions and with different backbones. Our approach significantly surpasses the other methods. Especially, compared to the baseline, our method achieves an improvement of 6.388 under the metric of PCKh@0.5 when the backbone is HRNet-W32 and the input resolution is 32 × 32. This further validates the effectiveness of MCDC.

5.4.2. Classification

In this section, we evaluate the performance of our MCDC module in low-resolution classification. The experiments are conducted on two datasets: CIFAR-10 [10] and TinyImageNet [9].

TinyImageNet. We replace the downsampling layer with MCDC, as shown in Table. 3. When we apply the MCDC on ResNet18 [40] on TinyImageNet, the performance of the network is significantly improved. Specifically, MCDC brings a 1.93% improvement to ResNet18, far exceeding that of Nystromformer [41] and WaveMix-128/7 [42]. It fully demonstrates the effectiveness of MCDC in low-resolution classification. In addition, when MCDC is applied to ConvNet-T, the performance improves by +38.09%. This is mainly because ConvNet-T has a relatively simple structure, which limits its ability to extract information, while MCDC fully exploits multi-granularity information, thereby significantly improving the model's performance.

CIFAR-10. Similarly, when we train and test our model on CIFAR-10, the performance of the network is also slightly improved. It outperforms the ResNet-50 baseline by an absolute margin of 0.26%. It shows that our method is versatile for low-resolution classification tasks. The effectiveness and generality of our proposed MCDC are validated.

Table 3
Result of our MCDC module in low-resolution image classification tasks.

Model	Dataset	Top-1 accuracy (%)↑
Nystromformer [41]	TinyImageNet	49.56
WaveMix-128/7 [42]	TinyImageNet	52.03
ConvNet-T(*baseline*) [43]	TinyImageNet	10.05
ConvNet-T-MCDC(*ours*)	TinyImageNet	**48.14**$_{(+38.09)}$
ResNet18(*baseline*) [40]	TinyImageNet	61.68
ResNet18-MCDC(*ours*)	TinyImageNet	**63.61**$_{(+1.93)}$
ResNet50(*baseline*) [44]	CIFAR-10	93.94
Stochastic Depth [45]	CIFAR-10	94.77
Prodpoly [46]	CIFAR-10	94.90
ResNet50-MCDC(*ours*)	CIFAR-10	**94.20**$_{(+0.26)}$

Table 4
Results of our MCDC module in low-resolution object detection tasks on COCO val set. * indicates that MCDC is only applied to the first downsampling layer to reduce training cost.

Method	Backbone	Image size	AP↑
YOLOv5-SPD-n [16]	YOLOv5	640 × 640	31.0
YOLOX-Nano [47]	YOLOv5	640 × 640	25.3
YOLOv5n(*baseline*) [47]	YOLOv5	640 × 640	28.0
YOLOv5n-MCDC(*ours*)	YOLOv5	640 × 640	**38.7**$_{(+10.7)}$
YOLOv5s(*baseline*) [47]	YOLOv5	640 × 640	37.4
YOLOv5s-MCDC*(*ours*)	YOLOv5	640 × 640	**38.8**$_{(+1.4)}$

Table 5
Ablation studies of each proposed component. Improv. = Improvement. It refers to the improvement of the method relative to the baseline (SimCC [26]). Indivi-improv. = Individual improvement. It refers to the improvement effect of the corresponding individual component. All experiments use HRNet-W32 as the backbone and take images with 32 × 32 resolution as input.

	FIP						
Method	SCT	DIE	CCA	CAGA	AP↑	Improv.	Indivi-improv.
SimCC [26]	–	–	–	–	29.8	–	–
Ours	✓	–	–	–	33.6	+3.8	+3.8
Ours	✓	✓	–	–	36.3	+6.5	+2.7
Ours	✓	✓	✓	–	36.8	+7.0	+0.5
Ours	✓	✓	✓	✓	37.5	+7.7	+0.7

5.4.3. Object detection

COCO. To demonstrate the generalizability of our approach, we further apply the module to low-resolution object detection tasks and conduct experiments on the COCO dataset. As shown in Table. 4, we replace the downsampling layers in YOLOv5n and YOLOv5s [47] with MCDC. These results demonstrate the effectiveness and versatility of our method in low-resolution object detection tasks. Specifically, YOLOv5n-MCDC achieves a 10.7 AP improvement over the baseline, while YOLOv5s-MCDC improves the AP by 1.4 compared to the baseline.

5.5. Ablation studies

Each component. In this subsection, we conduct ablation experiments to demonstrate the contribution of each proposed component. We replace the traditional downsampling convolutions with a stride of 2 in SimCC [26] with our proposed MCDC, while keeping the rest of the network structure unchanged. To verify the individual effect of the SCT unit, we only use the FIP submodule and replace the DIE block with a single convolution for dimension transformation. To evaluate the role of CAGA, we replace it with a multi-branch direct addition and compare the obtained results with those using CAGA. As shown in Table. 5, each of the proposed components contributes to the improvement of the model's performance. Compared to traditional down-sampling layers, the Space-Channel Transformation (SCT) unit yields a 3.8 AP improvement, indicating that it can better exploit fine-grained information that is typically lost during conventional down-sampling. Furthermore, the proposed Fine-grained Information Preservation (FIP) submodule brings a substantial gain of 6.5 AP over the baseline, suggesting its effectiveness in mitigating fine-grained information loss and learning more discriminative features. Building upon FIP, the Coarse-grained Context Aggregation (CCA) submodule provides an additional improvement of 0.5 AP, indicating that multi-granularity contextual information is effectively leveraged to refine model predictions. Finally, the Context-Aware Granularity Adapter (CAGA) further boosts performance by 0.7 AP, demonstrating its ability to better fuse multi-granularity features and generate higher-quality output representations. Overall, the ablation experiments

Table 6
Ablation studies of the number of non-strided convolutions. "Conv number" refers to the total count of non-strided convolutions in the FIP submodule.

Method	Conv number	Backbone	Input size	AP↑
SimCC(baseline) [26]	–	HRNet-W32	32×32	29.8
Ours	1	HRNet-W32	32 × 32	33.6
Ours	2	HRNet-W32	32 × 32	36.3
Ours	3	HRNet-W32	32 × 32	**36.6**

Table 7
Ablation studies of the multiple non-strided convolutions in different submodules. "Conv number" refers to the sum of non-strided convolutions.

Method	Conv number in FIP	Conv number in CCA	AP
SimCC(baseline) [26]	–	–	29.8
Ours	1	–	33.6
Ours	1	1	34.6
Ours	2	–	36.3
Ours	2	1	36.8
Ours	2	2	37.1

Table 8
Ablation studies for the Context-Aware Granularity Adapter (CAGA). "Conv number" refers to the sum of non-strided convolutions.

Method	Conv number in FIP	Conv number in CCA	CAGA	AP↑
Ours	1	1	×	34.6
Ours	1	1	✓	35.7
Ours	2	2	×	37.1
Ours	2	2	✓	37.5

confirm the complementary nature of each of our proposed components, and the combination of all proposed methods results in the best performance, with a significant 7.7 AP improvement over the baseline.

Non-strided conv number in the FIP submodule. To verify the impact of the number of non-strided convolutions in the FIP submodule on the model's performance, we conduct experiments with various numbers of non-strided convolutions. As shown in Table. 6, as the number of non-strided convolutions increases, the model's accuracy gradually improves, indicating that increasing the number of non-strided convolutions allows for further learning of low-resolution information. However, excessive non-strided convolutions also increase the model's computational cost. Therefore, we only conduct experiments using three non-strided convolutions.

Non-strided conv number in different submodules. Table. 7 shows the effect of the number of non-strided convolutions in different submodules. The results indicate that the multiple non-strided convolutions bring substantial improvement in both the FIP and CCA submodules. They effectively complement the SCT unit and the MRIE unit. They can be combined to further extract valuable information, significantly enhancing the model's performance. Considering the balance between performance and efficiency, we conducted experiments with only two convolutions. In practical applications, adjustments can be made according to the desired trade-off between the two.

CAGA. We perform ablation experiments on CAGA in different model structures. We implement different model structures by setting the varying number of non-stride convolutions in each submodule. As shown in Table. 8, relative to the direct additive fusion method, CAGA improves by 1.1 AP and 0.4 AP, respectively. It proves that the adapter can further extract effective feature information, illustrating better performance and robustness.

5.6. Results under normal resolution

We evaluate how our method performs on normal-resolution input. The results on the COCO validation set are shown in Table. 9. Although 128 × 128 is considered low-resolution, its visual quality is close to that of 256 × 192. Therefore, we also report the corresponding evaluation results. We can see that our MCDC module significantly outperforms the other methods. For example, our MCDC method outperforms SimCC [26] by +0.7 AP at the input size of 128 × 128. And under input resolution of 256 × 192, our method performs better again (*i.e.,* +0.3 AP and +1.7 AR). These results demonstrate that the proposed method maintains good generalization performance even under normal-resolution input conditions.

5.7. Evaluation of computational efficiency

To evaluate the computational efficiency of MCDC, we conduct a comparative experiment with standard down-sampling convolutions, and the results are shown in Table. 10. Compared with the baseline, MCDC introduces higher parameter count, GFLOPs, and inference time, but achieves a substantial performance improvement. The advantages of MCDC lie in its plug-and-play nature, design-free implementation, significant performance gains, and strong generalization capability. It is particularly suitable for general

Table 9
Comparison with SOTA methods on the COCO validation set under normal resolutions.

Method	Backbone	Input size	AP↑	AR↑
SimBase [25]	ResNet-50	128 × 128	55.4	63.3
Tokenpose [27]	Stem	128 × 128	57.6	64.9
HRNet [19]	HRNet-W48	128 × 128	63.3	70.5
SimCC(baseline) [26]	HRNet-W32	128 × 128	72.3	75.7
SimCC-MCDC(*ours*)	HRNet-W32	128 × 128	**73.0**	**76.1**
SimBase [25]	ResNet-50	256 × 192	68.5	74.8
Tokenpose [27]	Stem	256 × 192	69.9	75.8
HRNet [19]	HRNet-W48	256 × 192	73.1	78.7
SimCC(baseline) [26]	HRNet-W32	256 × 192	76.0	79.1
SimCC-MCDC(*ours*)	HRNet-W32	256 × 192	**76.3**	**80.8**

Table 10
Computational cost comparison with standard down-sampling. We replace the standard down-sampling in SimCC [26] with MCDC and adopt HRNet-W32 [19] as the backbone. The input resolution is set to 32 × 32.

Method	Params(M)↓	GFLOPs↓	Latency(ms)↓	AP↑
Standard Down-sampling	28.6	0.16	46.5	29.8
Ours	39.5	0.34	62.4	37.5

Table 11
Inference time breakdown of FIP, CCA, and CAGA. All experiments use HRNet-W32 as the backbone and take images with 32 × 32 resolution as input.

Method	FIP	CCA	CAGA	Latency(ms)↓
SimCC [26]	-	-	-	46.5
Ours	✓	–	-	56.5
Ours	✓	✓	-	59.8
Ours	✓	✓	✓	62.4

low-resolution scenarios that demand high accuracy while favoring simple solutions. Therefore, we believe that even a moderate increase in computational cost does not diminish the significance of MCDC in such cases.

5.8. Inference time analysis

To further investigate the computational efficiency, we conduct a runtime breakdown analysis of the FIP, CCA, and CAGA submodules. The inference-time results are reported in Table 11. The introduction of FIP, CCA, and CAGA all increases the overall computation time, resulting in additional latencies of 10.0 ms, 3.3 ms, and 2.6 ms, respectively.

This indicates that all three submodules introduce extra inference cost, with FIP and CAGA contributing most significantly to the overall computational overhead. The advantages of our method lie in performance optimization, generality, and plug-and-play capability, and a moderate increase in computational time is acceptable.

5.9. Memory overhead analysis

We compare the memory usage of standard down-sampling, SCT, FIP, CCA, and MCDC to analyze the memory overhead introduced by our method. As shown in Table. 12, MCDC achieves notable performance improvements at the cost of increased memory consumption. Specifically, SCT increases memory usage by 58.4 MB due to the expansion of channel dimensions. FIP introduces 92.9 MB, accounting for the largest portion of the memory increase in MCDC, while CCA alone adds 1.1 MB. Overall, MCDC trades additional memory usage for significant performance gains. Since low-resolution models typically have relatively low memory footprints, a moderate increase in memory consumption does not hinder practical deployment. Therefore, MCDC remains highly applicable in real-world scenarios.

5.10. Visualized results

Fig. 9 shows the visualization results of our method in benchmark scenes. Regardless of whether the input resolution is 64 × 64 or 128 × 128, the model can make accurate predictions. With 64 × 64 inputs, the model can still locate keypoints with relatively high precision and produce a clear skeletal structure. With 128 × 128 inputs, the model's performance is nearly comparable to that of humans. This demonstrates that our method can fully exploit and utilize the information contained in low-resolution images.

Table 12
Memory usage comparison among standard down-sampling, SCT, FIP, CCA, and MCDC. We implement the SCT approach by using FIP, where the DIE block is replaced with a single convolution. All methods are applied to SimCC [26].

Method	Backbone	Input size	Memory(M)↓
Standard Down-sampling	HRNet-W32	32 × 32	236.4
SCT	HRNet-W32	32 × 32	294.8
FIP	HRNet-W32	32 × 32	329.3
CCA	HRNet-W32	32 × 32	237.5
FIP + CCA	HRNet-W32	32 × 32	358.8
MCDC	HRNet-W32	32 × 32	364.1

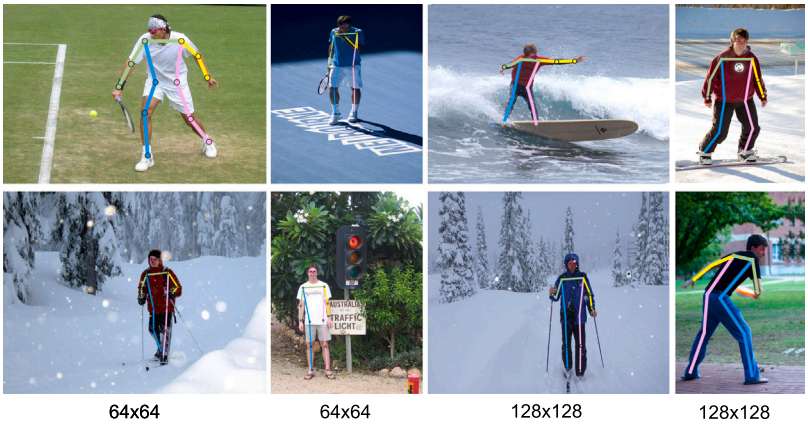

Fig. 9. Qualitative results of our method in benchmark scenarios. The images are obtained from the COCO validation set [7].

Fig. 10. Visualization comparison between our method **(bottom)** and the baseline **(top)** in real-world scenarios. "Red box" highlights the regions with significant differences. To simulate real-world scenarios, the model is trained on the COCO train dataset [7] and evaluated on the CrowdPose dataset [48], where images are closer to real-world conditions.

We visualized the comparison results between the baseline and our method for pose estimation in real-world scenarios. As shown in Fig. 10, our method achieves favorable performance under real-world conditions. Compared with the baseline, our predictions are more accurate, and in certain detailed regions, the results are closer to the ground truth. By retaining more effective information and fully leveraging contextual cues, our method improves the model's performance. These visualizations provide an intuitive demonstration of the practical value of MCDC.

6. Limitations and future work

As a novel downsampling convolution, MCDC provides a general solution for the problem of insufficient effective information learning in low-resolution vision tasks. However, the method relies on the consistency of training and testing resolutions, which is

unable to accommodate various resolutions at the same time. The limitation can be alleviated through multi-resolution adaptive training. Moreover, to adapt to the trend of large-scale models, exploring Transformer-based modules tailored for low-resolution vision represents an important direction for future research.

7. Conclusion

In this work, we propose a Multi-granularity Context-adaptive Downsampling Convolution (MCDC) for low-resolution images. The proposed convolution not only mitigates the significant information loss caused by traditional downsampling but also learns multi-granularity features and adaptively adjusts their contributions based on contextual information. MCDC consistently brings performance improvements across three different vision tasks and demonstrates strong generalization when applied to different models, backbones, datasets, and resolutions. Furthermore, this novel convolution introduces a new research paradigm for the design of low-resolution vision models, and the underlying modeling concept can also inspire computer vision tasks in complex scenarios, such as low-light, foggy, rainy, highly blurred, and dusty environments.

CRediT authorship contribution statement

Zejun Gu: Writing – review & editing, Writing – original draft, Visualization, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Zhong-Qiu Zhao:** Writing – original draft, Supervision, Funding acquisition, Conceptualization. **Hao Shen:** Writing – review & editing, Software, Project administration, Methodology, Investigation, Formal analysis. **Zhao Zhang:** Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis. **De-Shuang Huang:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62472137, 62502006, and 62072151), the Anhui Special Support Plan for High-Level Talents, the Guangxi Key Research and Development Program (Grant No. AB22035022), and the Scientific Research Foundation for High-level Talents of Anhui University of Science and Technology (Grant No. 2025yjrc0015).

Appendix A. Supplementary data

Supplementary data for this article can be found online at doi:10.1016/j.ins.2026.123223.

Data availability

The code is available at: https://github.com/guzejungithub/MCDC.

References

[1] M. Li, Y. Zhao, G. Gui, F. Zhang, B. Luo, C. Yang, et al., Object detection on low-resolution images with two-stage enhancement, Knowl.Based Syst. 299 (2024) 111985.
[2] S. Shin, J. Lee, J. Lee, Y. Yu, K. Lee, Teaching where to look: Attention similarity knowledge distillation for low resolution face recognition, in: European Conference on Computer Vision, Springer, 2022, pp. 631–647.
[3] L. Lo, B.K. Ruan, H.H. Shuai, W.H. Cheng, Modeling uncertainty for low-resolution facial expression recognition, IEEE Trans. Affect. Comput. 15 (1) (2023) 198–209.
[4] J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li, Q. Du, Superyolo: Super resolution assisted object detection in multimodal remote sensing imagery, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–15.
[5] C. Deng, M. Wang, L. Liu, Y. Liu, Y. Jiang, Extended feature pyramid network for small object detection, IEEE Trans. Multimedia 24 (2021) 1968–1979.
[6] H. Shi, X. Bai, C. Bai, Optimization of object detection network architecture for high-resolution remote sensing, Algorithms 18 (9) (2025) 539.
[7] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, et al., Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755.
[8] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, 2d human pose estimation: New benchmark and state of the art analysis, in: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition, 2014, pp. 3686–3693.
[9] Y. Le, X. Yang, Tiny imagenet visual recognition challenge, CS 231N 7 (7) (2015) 3.
[10] A. Krizhevsky, G. Hinton, Convolutional deep belief networks on cifar-10 40 (7) 2010, 1–9.
[11] M. Haris, G. Shakhnarovich, N. Ukita, Task-driven super resolution: Object detection in low-resolution images, in: International conference on neural information processing, Springer, 2021, pp. 387–395.
[12] F. Manessi, A. Rozza, M. Manzo, Dynamic graph convolutional networks, Pattern Recognit. 97 (2020) 107000.
[13] J. Cao, Y. Li, M. Sun, Y. Chen, D. Lischinski, D. Cohen-Or, et al., Do-conv: Depthwise over-parameterized convolutional layer, IEEE Trans. Image Process. 31 (2022) 3726–3736.
[14] Y. Xie, S. Li, C.T. Wu, Z. Lai, M. Su, A novel hypergraph convolution network for wafer defect patterns identification based on an unbalanced dataset, J. Intell. Manuf. 35 (2) (2024) 633–646.
[15] Z. Chen, B. Zhang, C. Du, W. Meng, A. Meng, A novel dynamic spatio-temporal graph convolutional network for wind speed interval prediction, Energy 294 (2024) 130930.

[16] R. Sunkara, T. Luo, No more strided convolutions or pooling: A new cnn building block for low-resolution images and small objects, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2022, pp. 443–459.
[17] G. Aloi, L. Bedogni, M. Felice di, V. Loscri, A. Molinaro, E. Natalizio, et al., Stem-net: an evolutionary network architecture for smart and sustainable cities, Trans. Emerg. Telecommun. Technol. 25 (1) (2014) 21–40.
[18] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[19] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 5693–5703.
[20] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, Eca-net: Efficient channel attention for deep convolutional neural networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11534–11542.
[21] R.R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, Int. J. Comput. Vis. 128 (2016) 336–359, https://api.semanticscholar.org/CorpusID:15019293.
[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, IEEE Conf. Comput. Vis. Pattern Recognit. (2015) 2818–2826, https://api.semanticscholar.org/CorpusID:206593880.
[23] G. Huang, Z. Liu, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2261–2269.
[24] C.H. Bennett, Logical reversibility of computation, IBM J. Res. Dev. 17 (6) (1973) 525–532.
[25] B. Xiao, H. Wu, Y. Wei, Simple baselines for human pose estimation and tracking, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 466–481.
[26] Y. Li, S. Yang, P. Liu, S. Zhang, Y. Wang, Z. Wang, et al., Simcc: A simple coordinate classification perspective for human pose estimation, in: European Conference on Computer Vision, Springer, 2022, pp. 89–106.
[27] Y. Li, S. Zhang, Z. Wang, S. Yang, W. Yang, S.T. Xia, et al., Tokenpose: Learning keypoint tokens for human pose estimation, in: Proceedings of the IEEE/CVF International conference on computer vision, 2021, pp. 11313–11322.
[28] C. Wang, F. Zhang, X. Zhu, S.S. Ge, Low-resolution human pose estimation, Pattern Recognit. 126 (2022) 108579.
[29] J. Li, S. Bian, A. Zeng, C. Wang, B. Pang, W. Liu, et al., Human pose regression with residual log-likelihood estimation, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 11025–11034.
[30] F. Zhang, X. Zhu, H. Dai, M. Ye, C. Zhu, Distribution-aware coordinate representation for human pose estimation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 7093–7102.
[31] Z. Geng, C. Wang, Y. Wei, Z. Liu, H. Li, H. Hu, Human pose as compositional tokens, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 660–671.
[32] S. Ye, Y. Zhang, J. Hu, L. Cao, S. Zhang, L. Shen, et al., Distilpose: Tokenized pose regression with heatmap distillation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2163–2172.
[33] S. Chen, Y. Zhang, S. Huang, R. Yi, K. Fan, R. Zhang, et al., Sdpose: Tokenized pose estimation via circulation-guide self-distillation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 1082–1090.
[34] Y. Xu, L. Zhao, C. Gong, G. Li, D. Wang, N. Wang, DynPose: largely improving the efficiency of human pose estimation by a simple dynamic framework, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2025.
[35] H. Wang, J. Liu, J. Tang, G. Wu, B. Xu, Y. Chou, et al., GTPT: group-based token pruning transformer for efficient human pose estimation, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, Cham, 2024, pp. 213–230.
[36] Z. Gu, Z.Q. Zhao, H. Ding, H. Shen, Z. Zhang, D.-S. Huang, Cross-domain knowledge distillation for low-resolution human pose estimation, arXiv:2405.11448, 2024, https://arxiv.org/abs/2405.11448.
[37] T. Jiang, P. Lu, L. Zhang, N. Ma, R. Han, C. Lyu, et al., Rtmpose: Real-time multi-person pose estimation based on mmpose, Comput. Res. Repos. abs/2303.07399 (2023).
[38] Z. Li, J. Ye, M. Song, Y. Huang, Z. Pan, Online knowledge distillation for efficient pose estimation, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 11740–11750.
[39] K. Li, S. Wang, X. Zhang, Y. Xu, W. Xu, Z. Tu, Pose recognition with cascade transformers, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 1944–1953.
[40] S. Targ, D. Almeida, K. Lyman, Resnet in resnet: Generalizing residual architectures, arXiv preprint arXiv:1603.08029, 2016.
[41] Y. Xiong, Z. Zeng, R. Chakraborty, M. Tan, G. Fung, Y. Li, et al., Nyströmformer: A nyström-based algorithm for approximating self-attention, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 14138–14148.
[42] P. Jeevan, A. Sethi, Wavemix: resource-efficient token mixing for images, arXiv preprint arXiv:2203.03689, 2022.
[43] Z. Liu, H. Mao, C.Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 11976–11986.
[44] B. Koonce, ResNet 50, in: Convolutional Neural Networks with Swift for TensorFlow: Image Recognition and Dataset Categorization, 2021, pp. 63–72.
[45] G. Huang, Y. Sun, Z. Liu, D. Sedra, K.Q. Weinberger, Deep networks with stochastic depth, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14, Springer, 2016, pp. 646–661.
[46] G.G. Chrysos, S. Moschoglou, G. Bouritsas, J. Deng, Y. Panagakis, S. Zafeiriou, Deep polynomial neural networks, IEEE Trans. Pattern Anal. Mach. Intell. 44 (8) (2021) 4021–4034.
[47] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, Y. Kwon, K. Michael, et al., ultralytics/yolov5: v7.0 – YOLOv5 SOTA realtime instance segmentation, Zenodo, 2022.
[48] J. Li, C. Wang, H. Zhu, Y. Mao, H.S. Fang, C. Lu, Crowdpose: Efficient crowded scenes pose estimation and a new benchmark, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 10863–10872.