Dynamic Unary Convolution in Transformers

Haoran Duan, Yang Long, Shidong Wang, Haofeng Zhang, Chris G. Willcocks and Ling Shao, *Fellow, IEEE*

Abstract—It is uncertain whether the power of transformer architectures can complement existing convolutional neural networks. A few recent attempts have combined convolution with transformer design through a range of structures in series, where the main contribution of this paper is to explore a parallel design approach. While previous transformed-based approaches need to segment the image into patch-wise tokens, we observe that the multi-head self-attention conducted on convolutional features is mainly sensitive to global correlations and that the performance degrades when these correlations are not exhibited. We propose two parallel modules along with multi-head self-attention to enhance the transformer. For local information, a dynamic local enhancement module leverages convolution to dynamically and explicitly enhance positive local patches and suppress the response to less informative ones. For mid-level structure, a novel unary co-occurrence excitation module utilizes convolution to actively search the local co-occurrence between patches. The parallel-designed Dynamic Unary Convolution in Transformer (DUCT) blocks are aggregated into a deep architecture, which is comprehensively evaluated across essential computer vision tasks in image-based classification, segmentation, retrieval and density estimation. Both qualitative and quantitative results show our parallel convolutional-transformer approach with dynamic and unary convolution outperforms existing series-designed structures.

Index Terms—Computer Vision, Transformer, Dynamic, Unary, Attention, Convolution

1 Introduction

BACKBONE deep model design has become the essential computer vision task [1]. Embracing the power of high-performance computing and rich visual contents from online platforms, the current leading paradigm of computer vision aims to pre-train a large-scale, multi-task, multi-modality model that can be transferred to downstream tasks. While Convolution-based deep neural Network (ConvNet) architectures have established a leading position in key computer vision tasks, e.g., image detection, classification, segmentation, the community has been seeking for multi-modal solutions since the last decade. An inevitable and essential topic is about sequential-modeling which has natural applications in videos, free-texts, audios, and many other signals of wearable devices. The traditional RNN-based paradigm was challenged by the transformer-based neural Network (TransNet) architecture [2] which has soon become a dominant approach. Recent research has shown that the TransNet architecture can even outperform ConvNet on pure vision tasks [3]. Debate has focused on whether the vision and language tasks should be brought together, and the model paradigm should be unified in the new TransNet formula for better transition between multi-modal tasks.

ConvNets have natural advantages in visual tasks due to their spatial prior. The existing TransNet paradigm breaks visual data into local patch tokens. The natural 2D or 3D neighborhood dependence is broken into a 1D sequential order. Fully-connected attention with dense tokens is needed to capture the dependence, which makes TransNets suffer from poor scalability and flexibility. A few recent attempts introduce convolution to transformers and achieved promising results [4]. As it is shown in Fig. 1, most exiting deep architectures adopt residual connections. Layers between two residual connections can be regarded as a block. For simplicity, the figure does not include normalization, activation, pooling, etc. Most TransNet also adapts residuals as shown in Fig. 1 (B). A straightforward approach is to break the low-level vision tasks using ConvNets and apply TransNet onto the feature maps to process the high-level information. Dai et al. has proposed a CoatNet that consists of consecutive conv blocks followed by further transformer layers [5]. This paradigm is illustrated in 1 (C). Further attempts concentrate on introducing convolution to transformers as a hybrid block as shown in 1 (D). It is intuitive

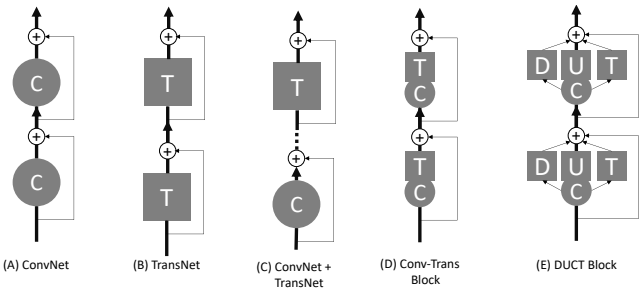


Fig. 1. Fig. 1. Comparison of existing convolution (A) and transformer (B) architecture designs with the proposed DUCT blocks. While previous work integrates convolution and transformer layers in a separate series (C), recent trends alternate transformer and convolution in a block-wise way (D). Our DUCT (E) is the proposed parallel structure combining a dynamic local enhancement module, a unary co-occurrence excitation module, and multi-head self-attention in a block-wise design.

- *Haoran Duan, Yang Long, and Chris G. Willcocks are with the Department of Computer Science, Durham University, UK. E-mail: haoran.duan@ieee.org; yang.long@ieee.org; christopher.g.willcocks@durham.ac.uk.*
- *Haofeng Zhang is with School of Computer Science and Engineering, Nanjing university of Science and Technology, China. E-mail: zhanghf@njust.edu.cn.*
- *Shidong Wang is with the School of engineering, Newcastle University, UK. E-mail: shidong.wang@ncl.ac.uk*
- *Ling Shao is with Terminus Group, China. E-mail: ling.shao@ieee.org*
- *Yang Long and Shidong Wang are the Corresponding authors.*

Manuscript submitted Feb 15, 2022;

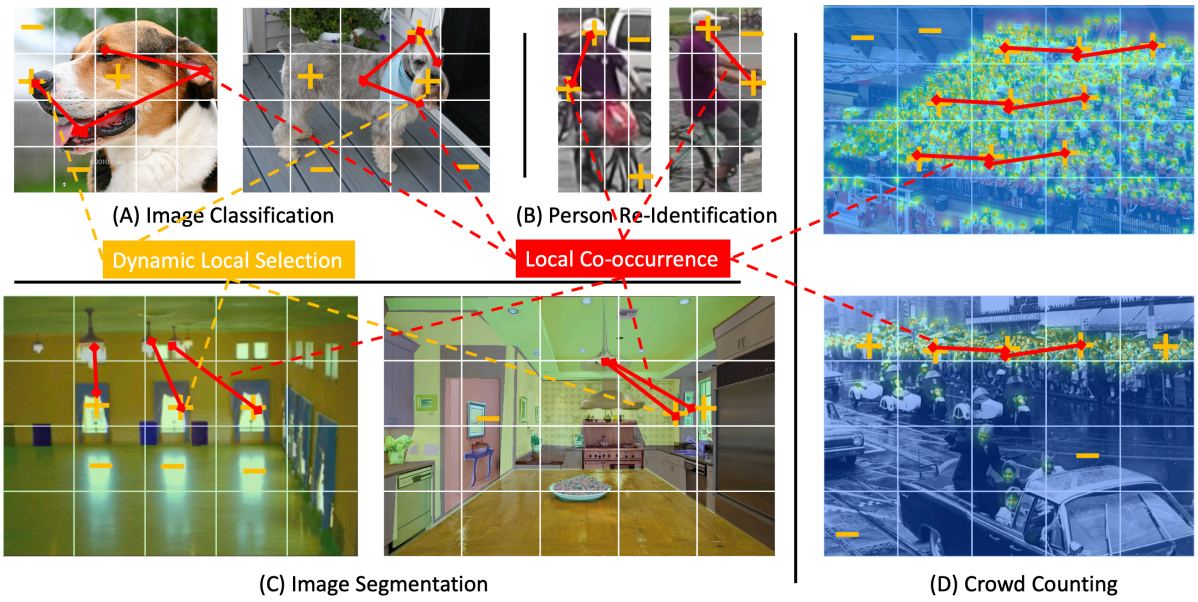


Fig. 2. Illustration of our proposed Dynamic Local Enhancement (DLE) and Unary Co-occurrence Excitation (UCE) in different computer vision tasks. DLE aims to assign weights to important local patches for convolution (in orange colour). UCE searches for unique co-occurrence between a local patch and others. Such co-occurrence at the feature-map level can achieve higher invariance. DLE, UCE and multi-head self-attention are combined to detect local, mid-level and global information in a complementary way.

to see that existing attempts for Conv-TransNet follow the design in a series order. The ConvNet before the transformer layer is interpreted as a tokenizer feature extraction of each image's local patch. The resulting feature maps are composed of the global information. Without further constraint or prior information, the following attention layer needs to compute the dependency between all of these feature maps. This property is particularly susceptible when the global structure of local patches is severely shifted, e.g. incomplete or occluded objects, or due to unseen viewing angles on the test domain.

In this paper, we explore parallel design to enhance the local and mid-level information as complementary to the global attention model, where the fusion of diverse features is hypothesized to boost the model performance on various scenarios. Our motivation is illustrated in Fig. 2. In most computer vision tasks, a few local patches are much more informative than others. Also, some patches can cause ambiguity and should be suppressed. When computing the global attention in an unnatural sequential order, the response from such local patch information is negated by other tokens, resulting in a blurring effect. Therefore, we first introduce a *Dynamic Local Enhancement* (DLE) module to achieve dynamic local selection, i.e., force the model to assign higher weights to important patches. For example, as illustrated in Fig. 2 (C), the reflection of window light on the floor can confuse the model and thus needs to be assigned lower weights. In contrast, the actual window region will receive high weights so that the convolution signal can be safely preserved, complementing the global attention.

The other novel idea comes from our observation of the local co-occurrence property. For example, the local patch of dog eyes often occurs within the (or nearby/alongside the) nose and mouth area. Such a correlation is very sensitive to shifting, e.g., view angles, occlusion, etc. Similarly, the crowd counting problem—shown in Fig. 2 (D)—is where crowded patches with similar heads are highly associated, and also associated with other crowds compared to non-crowded ones.

Taking advantage of the unnatural order of patch tokens, we can compute the affinity matrix of the patch tokens. Each row of the affinity matrix then represents the 1-to-n correlation for each of the local patch tokens. We develop a novel *Unary Co-occurrence Excitation* (UCE) module on the 1-to-n correlation vector. The key idea is that relative correlation can hold regardless of whether the positions are changed. As it is shown in Fig. 2 (C), the windows often co-occur with the lamp. In the two compared images, the patch locations of windows and lamps are very different. But the pair-wise or group-wise correlation score can hold for the tokens of windows and lamps regardless of the position shifting. It is named 'unary' because each patch token is only assigned with a single unique convolutional kernel, to search for similar score patterns. Also, it aims to search for invariant groups or pair-wise token correlations for each local patch. Such groups consist of several correlated patch tokens as a part of the global structure. Therefore, Unary Co-occurrence Excitation can provide mid-level information as a bridge between the local and global gaps. We summarize our main contributions as follows:

- The first attempt to integrate parallel structure within a hybrid Conv-Trans block.
- We introduce a dynamic local enhancement module to preserve highly informative local patch/token information.
- We propose a novel unary co-occurrence excitation module that searches for position-invariant local co-occurrence, achieved by convolution over group-wise correlation scores between patch tokens.

- The dynamic unary enhancement with Transformers is combined as a 3-channel block. And an adaptive patch merging process is designed to select diverse features and reduce redundancy. Finally, the DUCT deep architecture (consisting of aggregated DUCT blocks) is comprehensively evaluated in four essential computer vision tasks, i.e., image-based classification, segmentation, retrieval, and regression (density estimation). The proposed method outperforms existing Conv-Trans design in series with state-of-the-art results.

The proposed DUCT block and the parallel design aims to bring new theoretical insights and help future work build a large-scale architecture for extremely large datasets. Yet the evaluation on lots of large-scale datasets is out of the scope of this paper; our goal is to design a flexible and generic Conv-TransNet on different computer vision paradigms. The following paper is organized as follows. In section two, our literature review examines both pure transformers in vision and hybrid visual transformers. Our technical details and methodology are introduced in section three. In section four, we introduce the experimental design and results discussion of the model performance on four computer vision tasks. Theoretical statements are supported by both a qualitative and quantitative ablation study. Our work is summarized in the last section, where further work and potential impacts are discussed.

2 Related Work

Transformers benefit from the multi-head self-attention mechanism, and have become the prominent model in natural language processing (NLP) [6], allowing for information capture over different ranges. Recently, transformers and their variants have shown encouraging potential on computer vision tasks, and are considered to be alternative models to classical convolutional neural networks (CNNs). Here, we aim to summarize and discuss the recent development of pure transformer models and hybrid transformers.

2.1 Transformers in Computer Vision

Earlier research largely focuses on the differences between words and pixels, with various methods that apply the word embedding concept to image data. The interdependence among pixels is critical to be captured by the self-attention mechanism, hence Parmar et al. [7] conducted experiments for image generation tasks by applying self-attention for each query pixel instead of modeling them globally. With the pixel-wise channel-specific embedding and multi-head self-attention blocks, the proposed model achieved competitive performance on image recognition task without using extra convolutional blocks.

An alternative method to scale attention is to apply it in blocks of varying sizes; Cordonnier et al. [8] applied self-attention on top of 2x2 patches, although this does not generalize to large-scale vision tasks. Also, some works tried to increase the receptive field with different sizes of patches to flexibly handle larger resolution images. In recent years, transformers have emerged as the backbone that drives advances in image classification—traditionally dominated by CNNs. Dosowitzky et al. [3] proposed a Visual Transformer (ViT), which has a similar form to those used in NLP tasks. It performs well on image classification tasks directly applied to image patch sequences. The network adopts a similar approach to BERT's tokenization method, where a learnable embedding is applied to the sequence of embedding patches. The state of this embedding serves as the image representation. In addition, a learnable 1D positional embedding was added to the patch embedding to retain positional information. In most cases, ViT is pre-trained on large datasets such as the ImageNet and then fine-tuned for smaller downstream tasks. Beyond the ViT, a set of variants were proposed to improve the performance; mainly focusing on enhancing locality, improving self-attention performance and architecture design. For excavating local information in different scales and locations, TNT [9] further divides the patch used in ViT into multiple sub-patches, where a transformer-in-transformer architecture was developed to capture the relationship between such inner transformer blocks, and for patch-level information exchange an outer transformer block was developed. Swin Transformers [10], [11] conduct local attention within a window and introduce a shifted window partitioning approach for cross-window connections. Shuffle Transformer [12] further utilizes the spatial shuffle operation instead of shifted window partitioning to achieve cross-window connections. RegionViT [13] generates regional tokens and local tokens from an image, and each forward token receives global information via attention with regional tokens. However, for vision tasks, recent work suggests that the transformers focus more on the global features, where it is still uncertain if other lower levels of information are necessary. In this work, we observed that different levels of information are still critical in a vision recognition network. This motivated us to design a new paradigm of hybrid transformer network to enhance representation learning inside and across the different tokens.

2.2 Hybrid Vision Transformers

Many recent works have independently confirmed that transformers can be successfully applied to various vision tasks [14], [15], [16], [17], [18], [19], [20], as they are able to capture long-range dependencies in inputs. However, there are still gaps in performance when compared with traditional CNN-based networks. The performance of ViT is largely limited where the training data is inadequate, especially compared with that of state-of-the-art ConvNets. There have been some works that combine convolution with self-attention in recent years. Although it seems that tokenized embedding works well, the transformers still need to be enhanced to learn dense, repeatable patterns (e.g., textures and edges) which convolutions are significantly more efficient at learning. And these early frameworks generally require a significant increase in computing resources to outperform convolutional variants.

There has also been recent interest in combining CNNs with forms of self-attention. Existing research focuses on improving the capability of extracting local information. With the image domain-specific inductive biases, [4] proposed the CvT to combine CNNs and transformers to model both

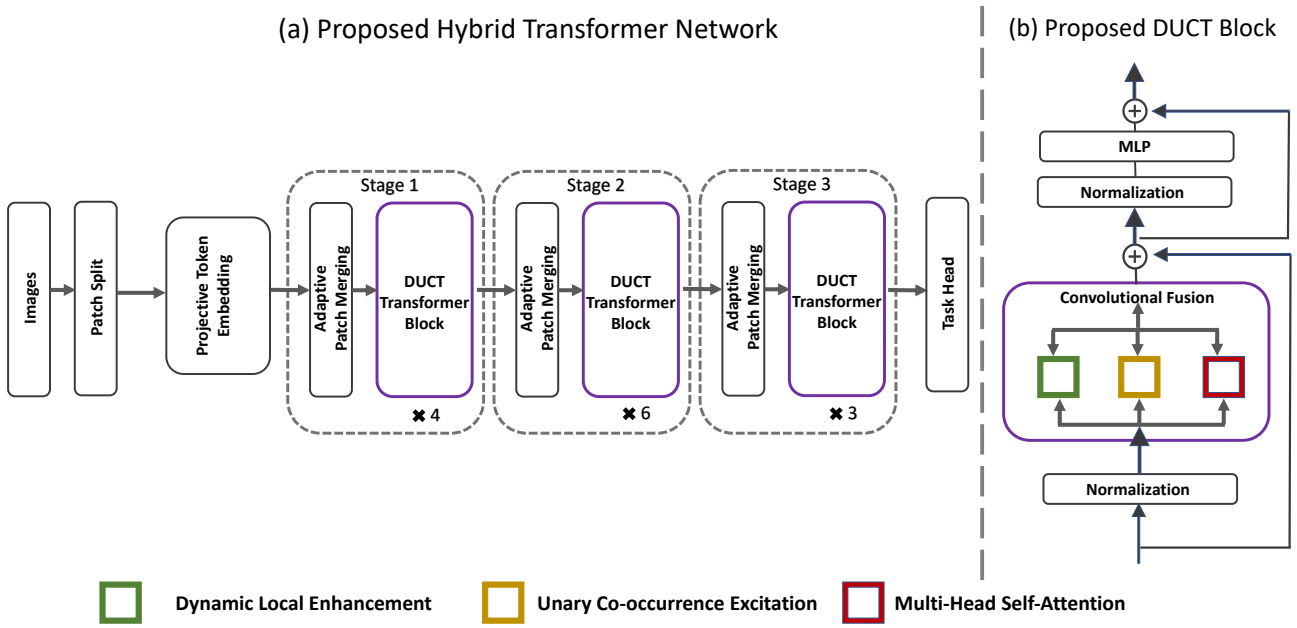


Fig. 3. (a) Architecture overview of the proposed hybrid transformer network DUCT. (b) The proposed hybrid transformer block for DUCT.

local and global dependencies for image classification in an efficient way. Their model consists of two novel structural changes. Firstly, they use convolutional projection modules to replace the existing position-wise linear projection for the attention operation. Secondly, they adopt a hierarchical multi-stage structure to support varied resolutions of 2D reshaped token maps, named convolutional token embeddings. Other works similarly analyzed the drawbacks of directly applying transformers from NLP on image tasks, such as [21], [22], [23]—focusing on either replacing or combining the feedforward network (FFN) with convolutional layers in each transformer module to better capture the correlation between neighboring tokens.

It's worth mentioning that there is also research into leveraging self-attention-style techniques to boost the performance of CNNs; [24] augment convolutions by concatenating convolutional feature maps with explicit self-attention. This additionally validates the benefits of combining both architectures. To enhance the model's awareness of global information, Wang et al., [25] proposed non-local operations as a family of building blocks that can capture long-range dependencies from sequences. Their approach achieves more accurate classification results for videos than 2D and 3D ConvNets, and is efficient in utilizing computational resources. The next year [26] introduced a geometric prior on the new local relation layer; the self-attention based layer extracts more representative compositional structures and adapts aggregation weights according to the spatial context. Most of the existing hybrid transformer paradigms attempted learning via series stream information, whereas investigating the parallel order of integrating information is an area that has not yet been well explored. This work aims to design a novel parallel hybrid transformer paradigm, where the goal is to enhance local, mid and high-level information; we hypothesize such diverse features are able to boost model performance in various vision tasks.

3 Methodology

The proposed approach aims to assemble off-the-shelf mainstream deep learning components in the most appropriate way to accomplish their mutual complementarity. Specifically, the details of each component in the proposed DUCT network will be outlined in the following sections; the Dynamic Local Enhancement module (DLE), the Unary Co-occurrence Excitation (UCE) module, the Multi-Head Vision Transformer (MHVT), together with discussions on the convolution operations accordingly.

3.1 Projection-enhanced Transformer

Vision Transformers [27] introduce a way to process input images in raster order, akin to word embeddings in transformers for NLP tasks. They use self-attention to substitute convolutional operations. Formally, given an input RGB image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ (H and W represent the height and width of the image respectively), the image is partitioned to generate N square patches (*a.k.a.*, tokens) where each patch has a spatial resolution of $\sqrt{P}\times\sqrt{P}$ and $N = \frac{H\times W}{\sqrt{P}\times\sqrt{P}}$. Note that there is no overlap between the adjacent patches. The resulting patches are then flattened and stacked to form $\mathbf{X} \in \mathbb{R}^{N\times P\cdot C}$, where C is the channel of each patch. According to the original Vision Transformer [27], $\mathbf{X}$ is linearly projected to a new embedding space of dimension $N \times C'$ to learn the global dependencies of tokens. However, the linear projection might potentially overlook some useful information because it can only reflect the local patterns represented by the split patches within their limited context.

Projective Token Enhancement To alleviate the impact of linear token embedding, a projective token enhancement module is proposed; a given input RGB image is split into non-overlapping patches $\mathbf{I'} \in \mathbb{R}^{(\frac{H}{4}\times\frac{W}{4})\times C'}$, which implies that the dimension of the flattened token is $C' = 4 \times 4 \times 3$.

The resulting patches are then fed into a projective token embedding module, similar to Swin Transformer [10]—composed of three linear mapping layers and normalization layers—but introduces an additional non-linear activation layer and the standard residual connection to generate the preliminary features of tokens, denoted as $\mathbf{X} \in \mathbb{R}^{N \times D}$.

These features, from the embedded tokens, allow for the low-level cues in each token to be preserved within the hierarchically designed embedding module, complimenting the convolution architecture. This early feature processing is distinct from previous work [3], [21], in that we use the transformer block to directly perform feature extraction on the embedded tokens.

Self-Attention Mechanism The use of self-attention [2], [27] allows for capturing global contextual dependencies present in the N entries of embedded features $\mathbf{X} \in \mathbb{R}^{N \times D}$. Specifically, $\mathbf{X}$ is encoded as the query $\mathbf{Q}$, the key $\mathbf{K}$, and the value matrix $\mathbf{V}$ of dimensions D_q, D_k, and D_v respectively. These act as the input of the self-attention layer. The output of the self-attention layer is a weighted sum of the values:

$$\text{Attention} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{D_k}}\right)\mathbf{V}. \tag{1}$$

In other words, the weight for each value is assigned by the scaled dot-product of each query and all keys.

Multi-Head Self-Attention Self-attention can be decomposed into multiple heads to support parallel and independent computation while considering the diversity of contextual information between patches and the aggregation of different representation subspaces. Specifically, with h as the number of attention heads and the learnable projection matrices $\mathbf{W}_i^Q$, $\mathbf{W}_i^K$, $\mathbf{W}_i^V$ and $\mathbf{W}^O$, Multi-Head Self-Attention (MHSA) is calculated in parallel:

$$\begin{aligned}\text{MHSA}\left(\mathbf{Q},\mathbf{K},\mathbf{V}\right) &= \text{Concat}\left(\text{head}_1,\ldots,\text{head}_h\right)\mathbf{W}^O \\ \text{where head}_i &= \text{Attention}\left(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V\right),\end{aligned} \tag{2}$$

and the output from the multi-head self-attention module is denoted $\mathbf{X}^{\tilde{\text{T}}} \in \mathbb{R}^{N \times D_v}$.

MHSA enables the Vision Transformer to capture global dependencies of the generated tokens without recursion. However, this comes at the expense of scalability in various computer vision tasks, which often require the proposed model to have a more targeted response to the task [28]. For example, as stated in [29], the transformer can only acquire effective local information by training on large-scale datasets (even larger than ImageNet). To this end, as shown in Fig. 3, we introduce a dynamic local enhancement and unary co-occurrence excitation module induced from standard convolutional operators to further enhance the expressiveness of features at different scales, thereby improving the modelling capability of the transformer for various tasks.

3.2 Dynamic Local Enhancement

Convolution-based deep learning models can extract local pixel information by means of using small filters, while the transformer blocks cannot explicitly model such fine-scale in a way that is scalable [10], [13], [28]. To enhance the ability of extracting the local features in each patch, a Dynamic Local Enhancement (DLE) module is presented to adaptively estimate a set of learnable convolution kernels that can independently model the relevant spatial information for an individual token (shown in Fig. 4). Given the availability of the generated token features $\mathbf{X} \in \mathbb{R}^{N \times D}$, the statistical information represented for each token can be reduced by averaging each row vector of $\mathbf{X}$:

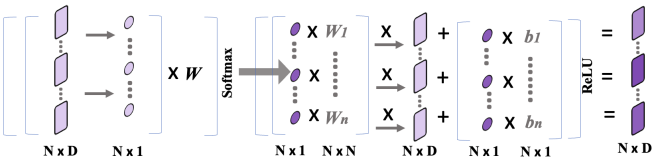


Fig. 4. The proposed Dynamic Local Enhancement (DLE) module. Given the token features, it first summarizes the average response, which is transformed to be the attention score. Then the attention score is used to calculate the dynamic convolution kernel for the dynamic local enhancement function.

$$\mathbf{x}_n^S = \frac{1}{D}\sum_{d=1}^{D}\mathbf{x}_n\left(1,d\right), \tag{3}$$

where $\mathbf{x}_n^S$ denotes the output of n_{th} row vector of $\mathbf{X}$. The overall representation $\mathbf{X}^S \in \mathbb{R}^{N \times 1}$ can be formed by stacking the averaged outputs of N row vectors, which is stable to the variations exhibited in the original features $\mathbf{X}$.

The values in $\mathbf{X}$ that summarize the average responses for all tokens are transformed to a set of attention scores by:

$$\mathbf{G}^S = \text{Softmax}\left(\mathbf{W}_2\left(\tau\left(\mathbf{W}_1\mathbf{X}^S\right)\right)\right), \tag{4}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are two weighting parameters, and τ is the ReLU activation function. Based on obtained attention score $\mathbf{G}^S \in \mathbb{R}^{N \times 1}$, we dynamically estimate the learnable kernels in order to enhance the variability of such attentive responses, where:

$$\mathbf{G}^{\mathbf{w}} = \sum_{n=1}^{N}\mathbf{G}_n^s\tilde{\mathbf{W}}_n, \quad \mathbf{G}^{\mathbf{b}} = \sum_{n=1}^{N}\mathbf{G}_n^s\tilde{\mathbf{b}}_n, \tag{5}$$

and $\tilde{\mathbf{W}} \in \mathbb{R}^{N \times N}$, $\tilde{\mathbf{b}} \in \mathbb{R}^{N \times 1}$ are the learnable weights and biases respectively. These are used to dynamically estimate the convolution kernel for different tokens. Furthermore, $\sum_{k=1}^{K}\mathbf{G}_k^s = 1$ with $0 \leq \mathbf{G}_k^s \leq 1$.

Based on the aggregation matrices $\mathbf{G}^{\mathbf{w}}$ and $\mathbf{G}^{\mathbf{b}}$, the locally enhanced features are obtained by:

$$\mathbf{X}^{\tilde{\text{D}}} = \tau\left(\mathbf{G}^{\mathbf{w}}\mathbf{X} + \mathbf{G}^{\mathbf{b}}\right), \tag{6}$$

where τ is the ReLU activation function. $\mathbf{G}^{\mathbf{w}}$ and $\mathbf{G}^{\mathbf{b}}$ denote the matrix transformation of a 1D convolution. In contrast to traditional 1D convolution, the weights of the convolution are dynamically assigned by $\mathbf{G}^{\mathbf{w}}$ and $\mathbf{G}^{\mathbf{b}}$. The resulting features $\mathbf{X}^{\tilde{\text{D}}}$ are of shape $N \times D_d$. Under the premise of retaining the feature size of the input token, the proposed DLE module can greatly increase the sensitivity of the local informative features and expand the network's ability to capture diverse information.

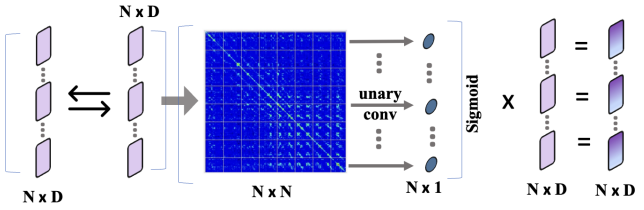


Fig. 5. The proposed Unary Co-occurrence Excitation (UCE) module. A correlation matrix is first calculated, and then it is transferred to the attention matrix by a unary convolution, which is used to enhance the 1-to-n correlation.

3.3 Unary Co-occurrence Excitation

Following the aforementioned observation of local co-occurrence, it is crucial for the model to learn the 1-to-n patterns to ensure that the correlations for different combinations of tokens are diverse regardless of changes in patch positions. To achieve this goal, we propose a novel Unary Co-occurrence Excitation (UCE) module, shown in Fig. 5. Considering the embedded features of tokens as $\mathbf{X} \in \mathbb{R}^{N\times D}$, the correlations between the tokens and channels are calculated as:

$$\mathbf{M} = \mathbf{X}\bar{\mathbf{I}}\mathbf{X}^{\top}, \tag{7}$$

where $\bar{\mathbf{I}} = \frac{1}{D}\left(\mathbf{I} - \frac{1}{D}\mathbf{1}\right)$ with an identity matrix $\mathbf{I} \in \mathbb{R}^{D\times D}$ and a matrix of ones $\mathbf{1} \in \mathbb{R}^{D\times D}$. The obtained correlation matrix $\mathbf{M} \in \mathbb{R}^{N\times N}$ can better reflect the pair-wise 1-to-n relationships among all tokens. More specifically, the diagonal entries of $\mathbf{M}$ represent the variances and the other entries represent the covariances between the tokens, meaning that the dependencies of the corresponding token with all other tokens are incorporated. Then, a unary convolution is proposed to effectively encode such 1-to-n relationships. Specifically, the matrix $\mathbf{M}$ is reshaped into $\mathbf{M} \in \mathbb{R}^{1\times N\times N}$, enabling convolution kernels $\mathbf{K}$ of size $1 \times N$ to be applied. This can be expressed by:

$$\tilde{\mathbf{M}} = \sigma\left(\mathbf{K}_{1,N,N} \cdot \mathbf{M}_{1,N,N}\right), \tag{8}$$

where σ is the sigmoid function, and $\cdot$ denotes the dot product indicating the convolution operation. The output of the UCE module $\mathbf{X}^{\tilde{\mathrm{U}}}$ is then just a product between the reshaped $\tilde{\mathbf{M}} \in \mathbb{R}^{N\times 1}$ and $\mathbf{X} \in \mathbb{R}^{N\times D}$:

$$\mathbf{X}^{\tilde{\mathrm{U}}} = \mathbf{X}\tilde{\mathbf{M}}. \tag{9}$$

Consequently, the resulting feature representations depicted above are concatenated to form a unified representation because they have identical dimensions. Formally, it is represented as:

$$\hat{\mathbf{X}} = \mathrm{Conv}\left(\mathrm{Concat}\left(\mathbf{X}^{\tilde{\mathrm{D}}}, \mathbf{X}^{\tilde{\mathrm{U}}}, \mathbf{X}^{\tilde{\mathrm{T}}}\right)\right), \tag{10}$$

where $\mathbf{X}^{\tilde{\mathrm{T}}}$ is the output from Multi-Head Self-Attention (MHSA). The concatenated representation has the shape $3N \times D$, which contains extensive useful information at different levels (low, mid and high levels), and the convolution operation is applied to filter the most valuable information while dropping the redundant information; the final output is then obtained, denoted as $\hat{\mathbf{X}} \in \mathbb{R}^{N\times D'}$.

3.4 Adaptive Patch Merging

Inspired by the recent work [4], [10], [30], a patch merging module is applied to combine the distinct feature representations. However, as reported in [31], merely applying regular grid-aware convolutional operations on the reshaped token sets [4], [30] may completely neglect the fact that different tokens usually contribute unequally, and also that tokens may have differing levels of interaction between each other. To handle these issues, motivated by deformable convolution [32], a group of offsets is introduced to effectively sample those informative tokens adaptively and then influence the process of merging tokens.

More formally, the unified features $\hat{\mathbf{X}} \in \mathbb{R}^{N\times D'}$ generated from Eq. (10) can be reshaped into $\hat{\mathbf{X}} \in \mathbb{R}^{1\times N\times D'}$. The standard convolution operation at the location $\mathbf{k}$ of each pixel can then be expressed as:

$$\bar{\mathbf{X}}\left(\mathbf{k}\right) = \sum_{\mathbf{k}_i \in [K\times K]} \mathbf{W}\left(\mathbf{k}_i\right) \cdot \hat{\mathbf{X}}\left(\mathbf{k} + \mathbf{k}_i\right), \tag{11}$$

where $\mathbf{k}_i$ enumerates the sampling locations in a convolution kernel (with size $K \times K$). A learnable offset $\Delta\mathbf{k}_i$ is then introduced into Eq.(11), yielding the adaptive patch merging scheme:

$$\bar{\mathbf{X}}\left(\mathbf{k}\right) = \sum_{\mathbf{k}_i \in [K\times K]} \mathbf{W}\left(\mathbf{k}_i\right) \cdot \hat{\mathbf{X}}\left(\mathbf{k} + \mathbf{k}_i + \Delta\mathbf{k}_i\right), \tag{12}$$

where the learnable offset $\Delta\mathbf{k}_i$ is estimated by an extra independent convolution layer. The output features are additionally transformed via a convolutional layer, a batch normalization layer, and then a GELU activation function. The dimensionality N is decreased to $\frac{1}{4}N$, and D' is increased accordingly to provide more channels/features information as in traditional convolutional neural networks. As the network depth increases, patch merging is used to reduce the number of tokens and control the channel dimension [3], [10] via adaptively integrating the informative patches, which allows for robust hierarchical representations giving the final output.

4 Experiments

Our experiments are conducted on four principal computer vision tasks including classification, segmentation, retrieval (person re-identification) and regression (crowd counting).

4.1 Model Configurations

Our model receives images of size 224×224 as input, which are initially partitioned into 4×4 patches. Then three linear embedding layers with normalization and residual connections are employed to preserve the subtle local information, and the output sequential token features are input to 3 hybrid transformer stages using the proposed DUCT blocks. The details of the DUCT blocks are described below:

- Stage 1: Patch size is 2 and the channel dimension is 128, the number of MHSA heads is 4, the number of transformer blocks is 4.
- Stage 2: Patch size is 2 and the channel dimension is 320, the number of MHSA heads is 6, the number of transformer blocks is 6.

TABLE 1
Comparisons with state-of-the-art methods on ImageNet-1K [29]

Method Type	Network	#Param.(M)	Image Size	FLOPs (G)	top-1 (%)
Convolution Neural Networks	ResNet-50 [33]	25	224 × 224	4.1	76.2
	ResNet-101 [33]	45	224 × 224	7.9	77.4
	ResNet-152 [33]	60	224 × 224	11	78.3
	RegNetY [34]	39	224 × 224	8	81.7
	EfficientNet [35]	19	380 × 380	4.2	82.9
Transformers	ViT-B/16 [3]	86	384 × 384	55.5	77.9
	ViT-L/16 [3]	307	384 × 384	191.1	76.5
	DeiT-S [36]	22	224 × 224	4.6	79.8
	DeiT-B [36]	86	224 × 224	17.6	81.8
	PVT-Small [30]	25	224 × 224	3.8	79.8
	PVT-Medium [30]	44	224 × 224	6.7	81.2
	T2T-ViTt-14 [37]	22	224 × 224	6.1	80.7
	T2T-ViTt-19 [37]	39	224 × 224	9.8	81.4
	TNT-S [9]	24	224 × 224	5.2	81.3
	TNT-B [9]	66	224 × 224	14.1	82.8
	Swin-T [10]	28	224 × 224	4.5	81.3
	Swin-S [10]	50	224 × 224	8.7	83.0
Convolution + Transformers	CvT-13 [4]	20	224 × 224	4.5	81.6
	CvT-21 [4]	32	224 × 224	7.1	82.5
	CoAtNet [5]	25	224 × 224	–	81.6
	MobileViT [38]	5.6	256 × 256	–	78.4
Ours (Hybrid Transformer)	DUCT_{224}	31	224 × 224	12	**83.1**
	DUCT_{384}	31	384 × 384	43.1	**84.7**

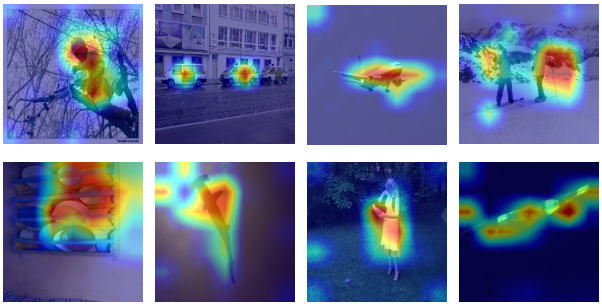


Fig. 6. Examples of class response maps from the output to the input on the ImageNet1K dataset.

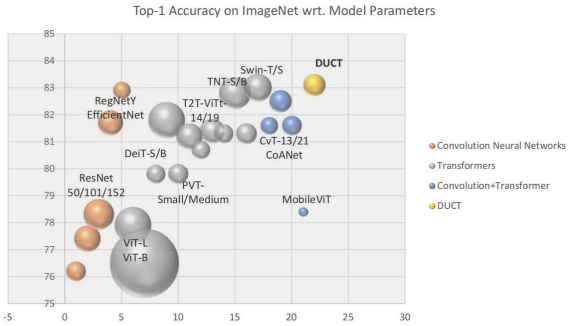


Fig. 7. The top-1 accuracy on ImageNet-1K [29] compared to other methods with respect to model parameters.

- Stage 3: Patch size is 2 and the channel dimension is 512, the number of MHSA heads is 8, the number of transformer blocks is 3.

4.2 Image Classification

The proposed DUCT is evaluated on five classification benchmark datasets, which are ImageNet-1K [29], CIFAR-10 [39], CIFAR-100 [39], Oxford Pet [40] and Oxford Flowers [41]. These experiments are set up as follows:

- ImageNet-1K [29] is a large-scale dataset which contains 1.28M training images and 50K validation images from 1,000 classes. The AdamW optimizer [42] with the cosine decay learning rate scheduler is used to optimize the network. The model is trained for 300 epochs with a batch size of 1024. The initial learning rate for training the entire model is set to 1e-3, and the weight decay is 0.005. The learning rate for adaptive patch merging is separately set to 1e-5. The top-1 accuracy and the computational costs are summarized for comparison.
- CIFAR-10 [39] has 50,000 training images and 10,000 testing images. CIFAR-100 [39] has 100 categories, each with 500 training images and 100 testing im-

ages per class. Oxford-Pets [40] has 37 categories, each with about 200 images per class of which 50 are for training, 50 for validation, and 100 for testing. Oxford-Flower [41] contains 102 flower categories and each class has around 40-258 images. We follow the previous work [4], [41] to split the train/validation/test sets. The model is fine-tuned on the model that was pretrained on ImageNet1K. The backbone is optimized using the SGD optimizer with a learning rate of 1e-4 and momentum of 0.9. It is trained for 200 epochs with input size 224×224 and batch size 256.

Table 1 shows the performance of DUCT on ImageNet1k compared with existing state-of-the-art methods based on CNNs, transformers and convolutional transformers. Also, a concise overview is shown in Fig. 7, based on the top-1 accuracy with respect to model parameters.

Compared with most state-of-the-art convolutional neural networks, the proposed DUCT achieves significantly higher top-1 accuracy. It also can be seen that DUCT achieves a better trade-off between accuracy and speed than existing CNN-based models. EfficientNet [35] and RegNetY [34] are the most recent mainstream convolution-based models; our model obtains competitive performance when compared with them. But they are built on neural architecture search [43] that usually requires a considerable amount of compute during the architecture search.

Recent advances in vision-based transformers have achieved great success in image recognition tasks and are comparative with CNN-based models. However, some transformer-based backbones require a considerable number of model parameters with only small improvements in results: ViT-L/16 [3] Swin [10] and DeiT-B [36]. Incorporating convolutions into transformers easily reaches an accuracy of 82%, where the proposed DUCT achieves competitive results over existing work in top-1 accuracy.

TABLE 2
Model performance on downstream tasks (* indicates that the experiments are conducted by ourselves.)

Method	CIFAR 10	CIFAR 100	Pets	Flowers 102
BiT-M [44]	98.91	92.17	94.46	99.30
ViT-B/16 [3]	98.95	91.67	94.43	99.38
ViT-L/16 [3]	99.16	93.44	94.73	99.61
ViT-H/16 [3]	99.27	93.82	94.82	99.51
EfficientNet [35]	98.90	91.7	95.40	98.8
RegNet* [3]	98.7	90.3	93.6	98.9
TNT-B [9]	99.1	91.1	95.0	99.0
CvT [4]	99.16	92.88	94.03	99.62
ours	**99.32**	**94.31**	94.76	99.55

Furthermore, to demystify the trustworthiness of the DUCT decision-making process, we leverage class activation maps to visualize the class responses of the entire DUCT model from output to input [45], [46]. Fig. 6 demonstrates that the proposed DUCT can highlight the accurate regions that are highly correlated with ground-truth semantic areas.

Moreover, to investigate the transferability of the pretrained DUCT model, we also fine-tune and evaluate it on several downstream datasets. Table 2 shows that the proposed DUCT can achieve reliable performance on downstream tasks.

TABLE 3
Model performance of semantic segmentation task on ADE20K dataset.

ADE20K Method	Backbone	mIoU	#param.
DLab.v3+ [50]	ResNet-101 [33]	44.1	63M
ACNet [51]	ResNet-101 [33]	45.9	38.5
OCRNet [52]	ResNet-101 [33]	45.3	56M
SemanticFPN [53]	ResNet101	38.8	48M
SemanticFPN [53]	PVT [30]	39.8	28M
SemanticFPN [53]	RegNet [30]	35.4	44M
SemanticFPN [53]	EfficientNet [30]	37.1	22M
SemanticFPN [53]	Swin [10]	41.5	32M
SemanticFPN [48]	**DUCT (ours)**	**42.1**	**37M**
UperNet [48]	ResNet-101 [33]	44.9	86M
UperNet [48]	DeiT [36]	44.0	52M
UperNet [48]	Swin [10]	46.1	60M
UperNet [48]	**DUCT (ours)**	**47.2**	**61M**

4.3 Image Segmentation

The widely-used semantic segmentation dataset ADE20K [47] is utilized to evaluate the effectiveness of the proposed DUCT backbone. ADE20K contains a total of 25k images that are labeled into 150 semantic categories; 20K of these images are used for training, 2K for validation and the remaining 3K for testing. While there may exist various other semantic segmentation frameworks, our goal is to fairly evaluate the proposed backbone performance. Hence, following the common practice [10], [30], we choose both the semantic-FPN [48], [49] and the UperNet [48] as the segmentation framework, and the model performance is measured by mIoU. AdamW [42] is used for optimization with a linear learning rate scheduler. The initial learning rate is set to 6e-5 with a weight decay of 0.01. The model is trained for 640k iterations with a batch size of 2.

Table 3 shows the semantic segmentation results on the ADE20K dataset. In comparing the proposed DUCT backbone with both convolution-based and transformer-based models, we find DUCT has superior performance in terms of mIoU, where it only incurs a slightly higher computational cost than others.

4.4 Density Estimation/Regression: Crowd Counting

To further reveal the generalizability of proposed DUCT, we further evaluate it on a density estimation/regression task, namely, crowd counting. Crowd density estimation aims to predict the density map of the number of target objects (e.g., people) in real-world images [54]. The experimental settings are the same as the recent transformer-based model [55] applied to the crowd counting. Benchmark datasets for evaluating the proposed model with DUCT blocks include ShanghaiTech_PartA [56], ShanghaiTech_Part B [56] and UCF_QNRF [57]. The model is trained for 2000 epochs and optimized with the AdamW [42] optimizer. The batch size is set to 4 with learning rate 1e-5. In addition, L2 regularization is adopted as commonly used to avoid overfitting.

As can be seen in Table 4, the transformer-based method [55] achieves competitive results compared to the latest work based on convolutional operations. While the proposed DUCT can further improve the estimation of crowd

TABLE 4
Model performance on crowd counting tasks.

Method	ST_Part A		ST_Part B		UCF_QNRF	
	MAE	MSE	MAE	MSE	MAE	MSE
PACNN [58]	62.4	102.0	7.6	11.8	-	-
S-DCNet [59]	58.3	95.0	6.7	10.7	104.4	176.1
DSSI-Net [60]	60.6	96.0	6.8	10.3	99.1	159.2
BL [61]	62.8	101.8	7.7	12.7	88.7	154.8
RPNet [62]	61.2	96.9	8.1	11.6	-	-
ASNet [63]	57.8	90.1	-	-	91.5	159.7
LibraNet [64]	55.9	97.1	7.3	11.3	88.1	143.7
AMRNet [65]	61.5	98.3	7.0	11.0	86.6	152.2
NoisyCC [66]	61.9	99.6	7.4	11.3	85.5	150.6
DM-Count [66]	59.7	95.7	7.4	11.8	85.6	148.3
GL [67]	61.3	95.4	7.3	11.7	84.3	147.5
SUA-Fully [68]	66.9	125.6	12.3	17.9	119.2	213.3
P2PNet [54]	52.7	85.1	6.3	9.9	85.3	154.5
BCCT [69]	53.1	82.2	7.3	11.3	83.8	143.4
CCTrans [55]	52.3	84.9	6.2	9.9	82.8	142.3
Ours	**52.1**	83.6	**6.1**	**8.6**	**82.1**	**141.5**

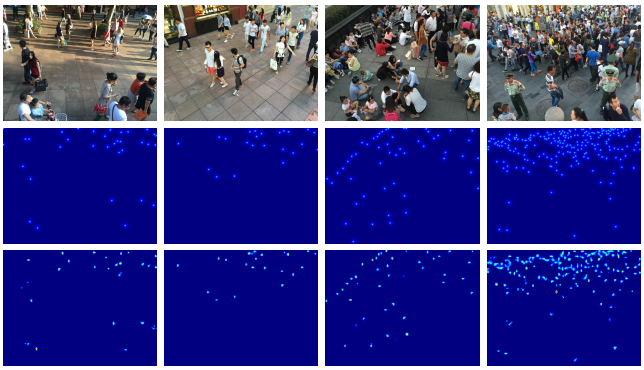


Fig. 8. Examples of estimated crowd density maps. From the first row to the last row, they represent the original images, the ground-truth density maps and the estimated density maps as predicted by DUCT.

density by properly combining the different feature representations. The qualitative visualizations of the estimated density maps are shown in Fig. 8.

4.5 Image Retrieval: Person Re-Identification

Image retrieval tasks involve searching for targets (e.g., images) from a gallery to match the query samples. Person re-identification, a prominent task in image retrieval, is considered for evaluating the efficacy of the proposed DUCT block. The experimental setup follows the recent work [70], where the framework built with the DUCT blocks is evaluated on Market1501 [71] and MSM17 [72] datasets.

Table 5 shows that recent transformer-based approaches perform slightly better than most CNN-based methods. The proposed backbone also achieves competitively with the latest transformer-based methods. Exemplar retrievals as obtained by the proposed backbone are shown in Fig. 9.

5 Further Discussion

In this section we discuss some key aspects of DUCT and its impact, primarily based on Fig. 10 and Table 6, where Fig. 10 shows the different levels of information (i.e., local, mid-level and global) learned by our model and Table 6 presents

TABLE 5
Model performance of person re-identification tasks.

Backbone	Method	Market1501		MSMT17	
		mAP	R1	mAP	R1
CNN	CBN [54]	77.3	91.3	42.9	72.8
	OSNet [73]	84.9	94.8	52.9	78.7
	MGN [74]	86.9	95.7	52.1	76.9
	RGA-SC [75]	88.4	96.1	57.5	80.3
	SAN [76]	88.0	96.1	55.7	79.2
	SCSN [77]	88.5	95.7	58.5	83.8
	ABDNet [78]	88.3	95.6	60.8	82.3
	PGFA [79]	76.8	91.2	-	-
	HOReID [80]	84.9	94.2	-	-
	ISP [81]	88.6	95.3	-	-
Transformer	TransReID(DeiT) [70]	88.1	94.9	65.5	83.5
	TransReID(ViT) [70]	88.8	95.0	66.6	84.6
	DUCT(ours)	**89.1**	95.1	**67.4**	**85.9**

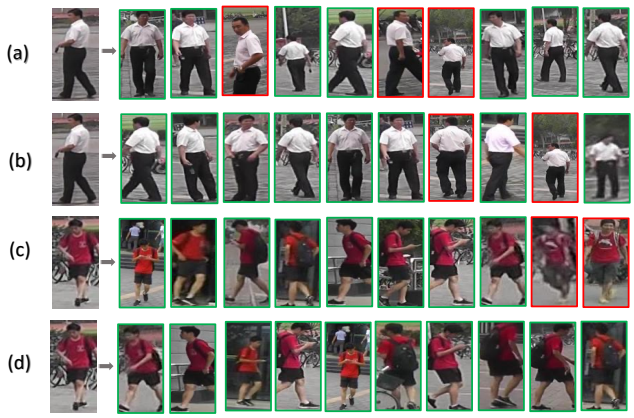


Fig. 9. Person retrieval samples from the Market1501 dataset. The first column is the query image, where others are retrieved images from the gallery, which is ranked according to the similarity scores. (a) and (c) are the results based on ViT-B/16. (b) and (d) are the results based on the proposed DUCT. GREEN indicates correctly matched samples and RED indicates mismatched samples.

the quantitative performance of each of the proposed components.

The impact of Dynamic Local Enhancement Since existing transformers are designed mainly for capturing long-range global information, one of our goals in this work is to enhance the local dependency. In Fig. 10 Row-2 blue curves, we observe that our proposed DLE module is able to highlight the potential response that global MHSA otherwise ignored. As the DLE is conducted based on summarizing each token, different local information is re-weighted from each token in a way that is complimentary for global information. Also, in Table 6, we can see that DLE quantitatively contributes to the final performance improvement.

The impact of Unary Co-occurrence Excitation The UCE module aims to leverage the mid-level information from groups of tokens. Rather than directly model different token information in a dense way, as MHSA, the mid-level information acts as the feature selector (Row-3 in Fig. 10) to assign the higher weights for correlated combinations of tokens—which can help discover finer information than dense MHSA from different token groups. The proposed

TABLE 6
Ablation study of the proposed components on different datasets and different tasks.

MHSA	APM	Dynamic	Unary	cifar100 (acc)	ImgNet (acc)	ADE20K (mIoU)	ST_A (MAE)	ST_B (MAE)	Market1501 (mAP)	MSMT17 (mAP)
✓				90.14	80.3	38.2	57.8	7.4	86.4	61.1
✓	✓			90.73	81.2	39.3	57.0	7.1	87.1	61.9
✓	✓	✓		91.83	81.9	40.7	53.4	6.6	87.6	63.5
✓	✓		✓	93.43	82.4	41.9	54.5	6.9	88.8	65.9
✓	✓	✓	✓	94.31	83.1	42.1	52.1	6.1	89.1	67.4

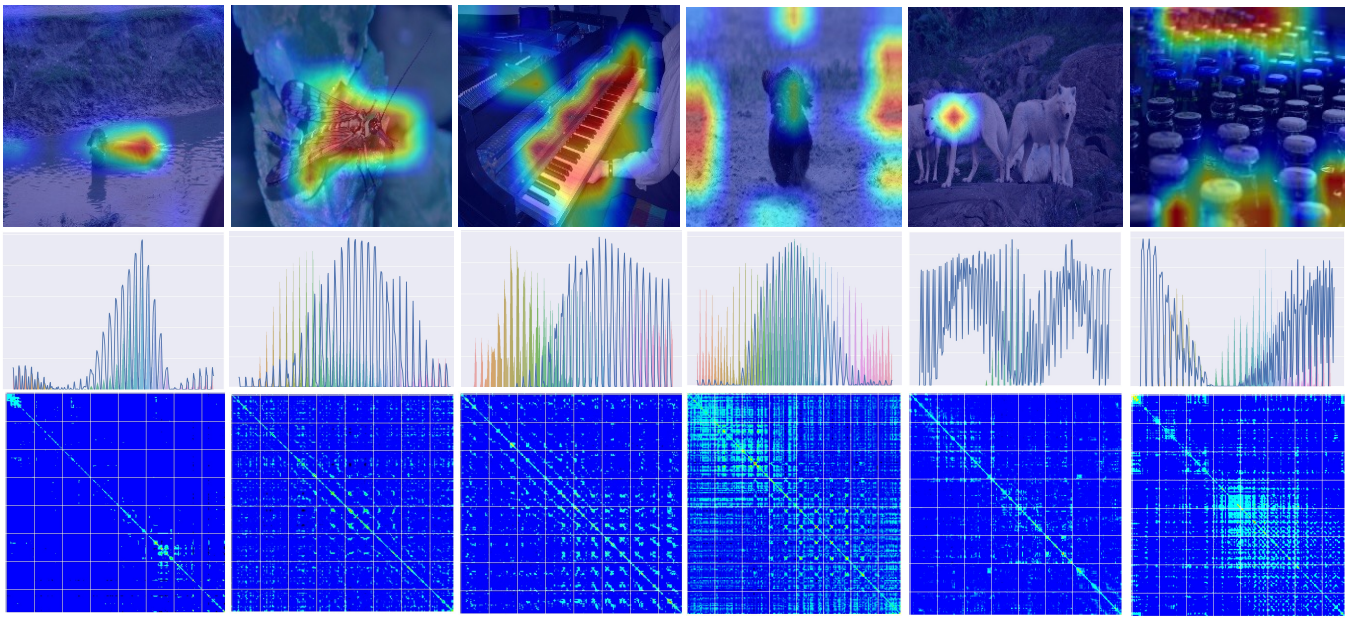


Fig. 10. Quantitative analysis of the response values of MHSA global attention, Dynamic Local Enhancement and Unary Co-occurrence Excitation. (Row-1) Visualization of the attention map. (Row-2) Comparison of Dynamic Local Enhancement (DLE) in the blue colour against global attention (MHSA) in the rainbow colour over local tokens. The x-axis is the tokens and the y-axis is the normalized attention response. (Row-3) Visualization of the correlation map in the Unary Co-occurrence Excitation module.

UCE also clearly contributes the final performance improvement as shown in Table 6.

Interaction/Limitation of different information levels The class response maps in Fig. 10 (Row-1) demonstrate that the proposed backbone can learn to precisely attend to the most relevant regions. Consistently, in the first column, the DLE and MHSA assigned similar weights on similar tokens, and the UCE unambiguously selects the most informative groups of tokens. A similar observation is also shown in the second column. In the third column, the DLE focuses on smaller regions of tokens since the black-white keys are the most distinct feature to recognize the piano, while the MHSA may further consider other parts of the piano. Also, the UCE still hold the capability to combine the accurate patterns/combination of different token groups. The last three columns show some failure examples, where we can see that there exists an obvious contradiction between DLE and MHSA, which has misguided the model's attention. Such failures are likely caused by a lack of controllable information selection. Since transformer architecture design is still a recent challenge and our main goal in this paper has been to design a novel parallel transformer architecture, we have directly concatenated the information using simple convolutional layers to select different information as learned by different components, whereas in the future, improvement could likely be made with improved selection of information from different blocks.

Design transformer for diverse vision tasks Existing transformer-based works mainly focus on popular vision tasks (e.g., classification, segmentation). In this paper, we also benchmarked our model on some other vision tasks like density estimation, image retrieval and downstream task transfer. The results demonstrate the potential of the DUCT framework and the importance of considering different levels of information. In previous years, CNNs have led progress in various vision tasks, with transformer-based models emerging as a breakthrough due to their expressivity and long-range information handling. Recent deep learning models are graduating towards a unification of various tasks and domains [82]. For the transformer family, this is both an opportunity and a challenge. The information required by different tasks and domains may be highly diverse, where designing a suitable universal model that is able to generalize across these diverse tasks and domains

remains an open problem. Our work proposed a parallel hybrid model, paving a novel approach to combine different levels of information into a single model, which could be a reference in future deep network backbone design.

6 Conclusion

In conclusion, we have proposed a hybrid transformer named DUCT. This is a parallel structure consisting of Dynamic Local Enhancement (DLE), Unary Co-occurrence Excitation (UCE), and a standard multi-head self-attention module, which together aim to learn the local, mid-level and global information. We found that DUCT outperforms the most recent state-of-the-art approaches on four essential computer vision tasks, i.e., image-based classification, segmentation, retrieval, and density estimation. This work paves a novel way to combine different levels of information, and the results reveal both the viability and validity of the approach. In the future, it would be worth investigating a more controllable selection of different levels of features (e.g., local and global) encoded in a hybrid transformer along with more in-depth theoretical analysis.

7 Acknowledgement

The authors would like to thank Bing Zhai for their helpful discussions. Ling Shao is partially supported by the National Natural Science Foundation of China (grant no. 61929104). Yang Long is supported by the U.K. Medical Research Council (MRC) Innovation Fellowship under Grant MR/S003916/2.

References

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020.
[4] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," pp. 22–31, 2021.
[5] Z. Dai, H. Liu, Q. Le, and M. Tan, "Coatnet: Marrying convolution and attention for all data sizes," *Advances in Neural Information Processing Systems*, vol. 34, 2021.
[6] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, pp. 1–67, 2020.
[7] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4055–4064.
[8] J.-B. Cordonnier, A. Loukas, and M. Jaggi, "On the relationship between self-attention and convolutional layers," no. CONF, 2020.
[9] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15 908–15 919, 2021.
[10] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," pp. 10 012–10 022, 2021.
[11] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "Cswin transformer: A general vision transformer backbone with cross-shaped windows," pp. 12 124–12 134, 2022.
[12] Z. Huang, Y. Ben, G. Luo, P. Cheng, G. Yu, and B. Fu, "Shuffle transformer: Rethinking spatial shuffle for vision transformer," *arXiv preprint arXiv:2106.03650*, 2021.
[13] C.-F. Chen, R. Panda, and Q. Fan, "Regionvit: Regional-to-local attention for vision transformers," 2021.
[14] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM computing surveys (CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022.
[15] A. A. Aleissaee, A. Kumar, R. M. Anwer, S. Khan, H. Cholakkal, G.-S. Xia *et al.*, "Transformers in remote sensing: A survey," *arXiv preprint arXiv:2209.01206*, 2022.
[16] J. Lahoud, J. Cao, F. S. Khan, H. Cholakkal, R. M. Anwer, S. Khan, and M.-H. Yang, "3d vision with transformers: A survey," *arXiv preprint arXiv:2208.04309*, 2022.
[17] M. Sultana, M. Naseer, M. H. Khan, S. Khan, and F. S. Khan, "Self-distilled vision transformer for domain generalization," *arXiv preprint arXiv:2207.12392*, 2022.
[18] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 519–16 529.
[19] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5728–5739.
[20] T. Sangam, I. R. Dave, W. Sultani, and M. Shah, "Transvisdrone: Spatio-temporal transformer for vision-based drone-to-drone detection in aerial videos," *arXiv preprint arXiv:2210.08423*, 2022.
[21] K. Yuan, S. Guo, Z. Liu, A. Zhou, F. Yu, and W. Wu, "Incorporating convolution designs into visual transformers," pp. 579–588, 2021.
[22] Y. Li, K. Zhang, J. Cao, R. Timofte, and L. Van Gool, "Localvit: Bringing locality to vision transformers," 2021.
[23] J. Guo, K. Han, H. Wu, Y. Tang, X. Chen, Y. Wang, and C. Xu, "Cmt: Convolutional neural networks meet vision transformers," pp. 12 175–12 185, 2022.
[24] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3286–3295.
[25] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.
[26] H. Hu, Z. Zhang, Z. Xie, and S. Lin, "Local relation networks for image recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3464–3473.
[27] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Computing Surveys (CSUR)*, 2021.
[28] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" *Advances in Neural Information Processing Systems*, vol. 34, 2021.
[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.
[30] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 568–578.
[31] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016, pp. 4905–4913.
[32] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.
[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[34] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 428–10 436.
[35] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[36] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 347–10 357.

[37] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," pp. 558–567, 2021.

[38] S. Mehta and M. Rastegari, "Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer," 2021.

[39] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[40] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3498–3505.

[41] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*. IEEE, 2008, pp. 722–729.

[42] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2018.

[43] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," *The Journal of Machine Learning Research*, vol. 20, no. 1, pp. 1997–2017, 2019.

[44] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big transfer (bit): General visual representation learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 2020, pp. 491–507.

[45] H. Chefer, S. Gur, and L. Wolf, "Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 397–406.

[46] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[47] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," *International Journal of Computer Vision*, vol. 127, no. 3, pp. 302–321, 2019.

[48] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 418–434.

[49] M. Contributors, "Mmsegmentation: Openmmlab semantic segmentation toolbox and benchmark," 2020.

[50] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[51] J. Fu, J. Liu, Y. Wang, Y. Li, Y. Bao, J. Tang, and H. Lu, "Adaptive context network for scene parsing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6748–6757.

[52] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 2020, pp. 173–190.

[53] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6399–6408.

[54] Q. Song, C. Wang, Z. Jiang, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and Y. Wu, "Rethinking counting and localization in crowds: A purely point-based framework," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3365–3374.

[55] Y. Tian, X. Chu, and H. Wang, "Cctrans: Simplifying and improving crowd counting with transformer," *arXiv preprint arXiv:2109.14483*, 2021.

[56] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 589–597.

[57] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, and M. Shah, "Composition loss for counting, density map estimation and localization in dense crowds," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 532–546.

[58] M. Shi, Z. Yang, C. Xu, and Q. Chen, "Revisiting perspective information for efficient crowd counting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7279–7288.

[59] H. Xiong, H. Lu, C. Liu, L. Liu, Z. Cao, and C. Shen, "From open set to closed set: Counting objects by spatial divide-and-conquer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8362–8371.

[60] L. Liu, Z. Qiu, G. Li, S. Liu, W. Ouyang, and L. Lin, "Crowd counting with deep structured scale integration network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1774–1783.

[61] Z. Ma, X. Wei, X. Hong, and Y. Gong, "Bayesian loss for crowd count estimation with point supervision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6142–6151.

[62] Y. Yang, G. Li, Z. Wu, L. Su, Q. Huang, and N. Sebe, "Reverse perspective network for perspective-aware object counting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4374–4383.

[63] X. Jiang, L. Zhang, M. Xu, T. Zhang, P. Lv, B. Zhou, X. Yang, and Y. Pang, "Attention scaling for crowd counting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4706–4715.

[64] L. Liu, H. Lu, H. Zou, H. Xiong, Z. Cao, and C. Shen, "Weighing counts: Sequential crowd counting by reinforcement learning," in *European Conference on Computer Vision*. Springer, 2020, pp. 164–181.

[65] X. Liu, J. Yang, W. Ding, T. Wang, Z. Wang, and J. Xiong, "Adaptive mixture regression network with local counting map for crowd counting," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*. Springer, 2020, pp. 241–257.

[66] J. Wan and A. Chan, "Modeling noisy annotations for crowd counting," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[67] J. Wan, Z. Liu, and A. B. Chan, "A generalized loss function for crowd counting and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1974–1983.

[68] Y. Meng, H. Zhang, Y. Zhao, X. Yang, X. Qian, X. Huang, and Y. Zheng, "Spatial uncertainty-aware semi-supervised crowd counting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 549–15 559.

[69] G. Sun, Y. Liu, T. Probst, D. P. Paudel, N. Popovic, and L. Van Gool, "Boosting crowd counting with transformers," *arXiv preprint arXiv:2105.10926*, 2021.

[70] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 013–15 022.

[71] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Computer Vision, IEEE International Conference on*, 2015.

[72] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 79–88.

[73] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3702–3712.

[74] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 274–282.

[75] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-aware global attention for person re-identification," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2020, pp. 3186–3195.

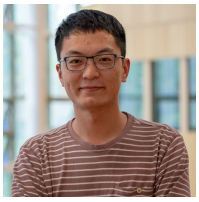
[76] X. Jin, C. Lan, W. Zeng, G. Wei, and Z. Chen, "Semantics-aligned representation learning for person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 173–11 180.

[77] X. Chen, C. Fu, Y. Zhao, F. Zheng, J. Song, R. Ji, and Y. Yang, "Salience-guided cascaded suppression network for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3300–3310.

[78] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang, "Abd-net: Attentive but diverse person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8351–8361.
[79] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 542–551.
[80] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, and J. Sun, "High-order information matters: Learning relation and topology for occluded person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6449–6458.
[81] K. Zhu, H. Guo, Z. Liu, M. Tang, and J. Wang, "Identity-guided human semantic parsing for person re-identification," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. Springer, 2020, pp. 346–363.
[82] R. Girdhar, M. Singh, N. Ravi, L. van der Maaten, A. Joulin, and I. Misra, "Omnivore: A single model for many visual modalities," pp. 16 102–16 112, 2022.

Haoran Duan (Student member, IEEE) received a Distinction M.S. degree in Data Science from Newcastle University, UK, in 2019. After that, he was a research student in OpenLab, Newcastle University, UK, and he is also a research associate at School of Computing, Newcastle University working on deep learning applications. He is currently pursuing a PhD degree in the Department of Computer Science, Durham University. His current research interests focus on the applications/theories of deep learning. He is the reviewer of CVPR, ECCV, AAAI, BMVC, TCSVT, TMM and Ubicomp.

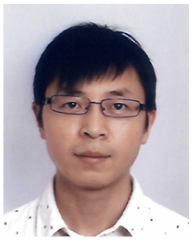
Yang Long is an Assistant Professor in the Department of Computer Science, Durham University. He is also an MRC Innovation Fellow aiming to design scalable AI solutions for large-scale healthcare applications. His research background is in the highly interdisciplinary field of Computer Vision and Machine Learning. While he is passionate about unveiling the black-box of AI brain and transferring the knowledge to seek Scalable, Interactable, Interpretable, and sustainable solutions for other disciplinary researches, e.g. physical activity, mental health, design, education, security, and geoengineering. He has authored/coauthored 30+ top-tier papers in refereed journals/conferences such as IEEE TPAMI, TIP, CVPR, AAAI, and ACM MM, and holds a patent and a Chinese National Grant.



Shidong Wang is a Research Associate at NEOLab, School of Engineering, Newcastle University. He received his PhD degree in 2021 from the School of Computing Sciences, University of East Anglia. His research spans a breadth of domains including computer vision, deep learning, remote sensing and environmental monitoring.

Chris G. Willcocks is an assistant professor in computer science at Durham University, where his interdisciplinary research focuses on generative models, medical image computing, computational biophysics, anomaly detection and machine reasoning. He teaches deep learning, reinforcement learning, and cyber security, and he regularly publishes in top-tier journal and conference papers in venues such as ICLR, PRX, ECCV, IEEE TPAMI, TMI, and TIFS.

Haofeng Zhang (Member, IEEE) received the B.Eng. and Ph.D. degrees from the School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing, China, in 2003 and 2007, respectively. From December 2016 to December 2017, he was an Academic Visitor at the University of East Anglia, Norwich, U.K. He is currently a Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests include computer vision and mobile robot.



Ling Shao (Fellow, IEEE) is the Chief Scientist of Terminus Group and the President of Terminus International. He was the founding CEO and Chief Scientist of the Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. His research interests include computer vision, deep learning, medical imaging and vision and language. He is a fellow of the IEEE, the IAPR, the BCS and the IET.